UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

DIVISIÓN DE AGRONOMIA

Estadística Básica y su Utilización Agropecuaria

POR:

SALVADOR GUERRERO LUNA

MONOGRAFIA

Presentada como Requisito Parcial para
Obtener el Titulo de:

Ingeniero Agrónomo Fitotecnista

Buenavista, Saltillo, Coahuila, México.

FEBRERO 2002

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO
DIVISION DE AGRONOMIA

Estadística Básica y su Utilización Agropecuaria

POR:
SALVADOR GUERRERO LUNA

MONOGRAFIA

Que somete a la consideración del H. Jurado examinador como requisito parcial para obtener el titulo de:

Ingeniero Agrónomo Fitotecnista

APROBADA

Asesor principal

_____

MC. Jaime M Rodríguez Del Ángel

Asesor **Asesor**

_____ _____

MC Humberto Macias Hernández **ING. Jesús Macias Hernández**

El Coordinador de la División de Agronomía

_____

MC Reynaldo Alonso Velasco

Buenavista, Saltillo, Coahuila Febrero 2002

INDICE DE CONTENIDO

INTRODUCCION

El objetivo de estas notas intituladas como monografía, por cuestión de reglamentación académica, no es el establecer un tratado sobre la aplicación de la Estadística en los aspectos agropecuarios, debido a que estas posiblemente no contienen nada nuevo para las personas que en su diario hacer se dedican a la investigación o evaluación de fenómenos en el campo de las ciencias naturales y que obviamente poseen un mayor cúmulo de conocimientos y experiencias. Sin embargo por interés propio y el demostrado por algunos de mis compañeros de estudios para los que nos pareció interesante este tema, es que me decidí a compilar y organizar las siguientes notas.

En la actualidad la Estadística, forma parte de nuestra vida diaria, basta con que leamos la prensa escrita o cualquier medio electrónico para darnos cuenta que la información esta dada en términos de porcentajes, medias, varianzas, coeficientes de variación etc. No hace muchos meses tuvimos elecciones en nuestro país y observamos, como el muestreo estadístico de encuesta y la probabilidad jugaron un papel preponderante al grado tal de omitir en algunos casos el error, sin embargo, para nosotros como estudiantes, la palabra Estadística, Probabilidad o Diseños Experimentales siempre represento un tema reservado solo para las personas con una gran preparación matemática y con dedicación meramente científica. Durante los cursos formales de esta materia en la Universidad, nos dimos cuenta que en nuestra especialidad como en muchas la Estadística es una herramienta importante para el estudio y evaluación de los fenómenos dentro del método científico y que no necesariamente se requiere ser un especialista en la materia para hacer uso de la misma en la explicación de las respuestas, sino más bien, es importante el comprender la correcta aplicación de las técnicas estadísticas al estudio del fenómeno. Lo anterior es comprensible si consideramos la definición que Snedecor y Cochran, en su libro Métodos Estadísticos dan para Diseños Experimentales, ahí mencionan que el procedimiento Estadístico para la evaluación de respuestas debe estar

considerado de antemano al pretender estudiar correctamente el comportamiento de un fenómeno dentro del método científico.

Otro aspecto que también contribuye a que la estadística no sea una de las materias mas comprendidas y apreciadas durante nuestros estudios Universitarios, se refiere al volumen de datos, simbología y ecuaciones que por normatividad deberán utilizarse; Por principio de cuentas no manejamos los programas estadísticos contenidos en las calculadoras portátiles, por demás decir los computacionales. Con respeto dicho, algunos maestros pretenden que de memoria tengamos presente la simbología y formularios que deberán utilizarse en cada uno de los procedimientos, además de que pocos son los ejemplos que contemplan algún aspecto relacionado con los procesos agropecuarios, debido principalmente a que pocos son los estadísticos, que tienen un perfil profesional relacionado con la Agronomía.

En la presente monografía, se contemplan temas estadísticos básicos que normalmente son incluidos en los cursos formales. Las medidas de Centralidad y Dispersión.- La media y la varianza, son medidas distintivas de una muestra y definen en primera instancia el comportamiento, en función de un parámetro de interés, así mismo es importante hacer notar la aplicación de otro tipo de medidas especiales, como la Moda, Mediana, Desviación media, Coeficiente de variación etc. en situaciones donde se desea identificar mediante algún atributo la muestra. Probabilidad.- En principio la probabilidad es un elemento indispensable en el proceso de inferencia estadística y es determinante en las pruebas de hipótesis e intervalos de confianza sobre los estadísticos descriptivos, resulta por demás importante cuando se trabaja con funciones de probabilidad de variables aleatorias discretas y continuas, tales como la Binomial, Hipergeometrica y Poisson en las discretas y Normal, Ji cuadrada, T de Student y F en las continuas.

Los estadísticos de prueba, de hipótesis que consideran variables aleatorias discretas pueden ser definidos a través de las distribuciones de probabilidad ya

mencionadas y dependerá del tamaño de la muestra, sesgo y preescisión la utilización especifica de cada una de ellas. Las pruebas de hipótesis sobre la media para variables aleatorias continuas, pueden ser llevadas a cabo mediante la utilización de distribuciones como la Normal y t de Student, así mismo la determinación de intervalos de confianza y comparación entre medias muéstrales se puede definir con la utilización de estas distribuciones. Cuando se habla de dispersión la distribución Ji cuadrada es de gran ayuda ya que la misma permite establecer pruebas de hipótesis sobre la varianza, intervalos de confianza, tablas de contingencia y pruebas de comportamiento a priori. Las pruebas de homogeneidad, Aditividad e independencia, de las observaciones de una muestra pueden ser estudiadas en función del planteamiento de hipótesis basadas en una distribución F de Fischer. Como se observa en general, la estadística básica ocupa un gran espacio en el proceso de inferencia dentro del estudio de los sucesos agropecuarios.

De los objetivos de esta monografía podríamos decir lo siguiente;

Mediante la recopilación, ordenamiento y presentación de los temas aquí expuestos, se obtendrá una experiencia que será útil en el ejercicio profesional.

Considerando que los ejemplos contenidos en los temas se refieren a aspectos agropecuarios, se tendrá un conocimiento más acorde respecto a la aplicación de las técnicas estadísticas y su interpretación en esta área de estudio.

Teniendo en cuenta lo estipulado en el REGLAMENTO ACADEMICO PARA ALUMNOS DE NIVEL LICENCIATURA DE LA U A A A N, articulo 85° fracción IV, obtener el Titulo Profesional de Ing. Agrónomo.

PROBABILIDAD

INTRODUCCIÓN

Históricamente la teoría de la probabilidad comenzó con el estudio de los juegos de azar tales como la ruleta y las cartas.

La probabilidad proporciona a quien toma una decisión un medio cuantitativo de expresar sus ideas sobre cada resultado.

¿Qué quieren decir, realmente, esas frases que se leen en los diarios tales como, la probabilidad de un brote de la enfermedad en el ganado de las *vacas locas* en México en los próximos años es de .5, la probabilidad de que el agro mexicano crezca en este sexenio es de .8?. Estas situaciones tienen la característica de no poder ser interpretadas en términos de frecuencias; no pueden ser repetidas, ni se repetirán . De este modo el significado de la palabra probabilidad no debe interpretarse como una alternativa a largo plazo. Sin embargo, se supone que las frases anteriores muestran un uso legítimo del concepto probabilidad. Al usar probabilidad en esta manera, se dice que está expresando el grado de credibilidad racional. Tales son consideradas como personales o subjetivas. Las persona asignarán probabilidades en base a su propia experiencia, antecedentes y conocimiento.

A continuación en este trabajo de investigación documental, se dará a conocer la teoría de la probabilidad, algunos de sus conceptos y ejemplos aplicados en actividades agrícolas y pecuarias para una mejor comprensión del tema.

TEORÍA DE LA PROBABILIDAD

DEFINICIÓN CLÁSICA DE PROBABILIDAD

Supóngase un suceso E, que de un total de n casos posibles, todos igualmente factibles, pueden presentarse en h de los casos. Entonces la probabilidad de aparición del suceso (llamada su *ocurrencia*) viene dada por :

$$p = P\{E\} = h / n$$

La probabilidad de no aparición (llamada su *no ocurrencia*) viene dada por ;

$$q= P\{ \text{no } E\} = (n – h) / n = 1 - h / n = 1- p = 1 - P\{E\}$$

Así, pues, $p + q = 1$ o $P\{E\} + P\{ \text{no } E\} = 1$
El suceso *no E a veces se denota por* $\bar{E}$ *ó* $\sim E$.

DEFINICIONES Y CONCEPTOS UTILIZADOS EN LA TEORÍA DE LA PROBABILIDAD.

ESPACIO MUESTRAL

Con cada experimento *E* del tipo que consideramos, definimos al espacio muestral como el conjunto de todos los resultados posibles de *E*, usualmente designamos este conjunto como *S*.

SUCESO

Un suceso *A, (*respecto a un espacio muestral particular asociado con un experimento *E)* es simplemente un conjunto de resultados posibles .
Se dice que dos suceso *A y B* son mutuamente excluyentes si no pueden ocurrir juntos, expresamos esto escribiendo $A \cap B = \emptyset$ es decir la intersección de *A y B* es el conjunto vacío.

FRECUENCIA RELATIVA

Fa = na / a se llama frecuencia relativa del suceso *A* en las *n* repeticiones de *E* . La frecuencia relativa tiene las siguientes propiedades

a) $0 \leq Fa \leq 1$

b) Fa = 1 Si y solo si *A* ocurre cada vez en las *n* repeticiones.

c) Fa = 0 Si y solo si *A* nunca ocurre en las *n* repeticiones.

d) Si *A* y *B* son dos suceso que se excluyen mutuamente y si Fa * b es la frecuencia relativa asociada al suceso *A* U *B* , entonces Fa U b = Fa + Fb

e) Fa basada en las N repeticiones del experimento y considerada para una función de n "converge" en cierto sentido probabilístico a *p(A)* cuando n - > 00

OBSERVACIÓN

Una de las características básicas del experimento es que no sabemos que resultado particular se obtendrá al realizar el mismo. En otras palabras si *A* es un suceso asociado con un experimento no podemos indicar con certeza que *A* ocurrirá o no. Por lo tanto llega a ser muy importante tratar de asociar un número con el suceso *A* que medirá de alguna manera, la probabilidad de que el suceso *A* ocurra.

VARIABLES ALEATORIAS

Una variable aleatoria es intuitivamente un método de asignar números o vectores de números a los resultados de un experimento .

Sea un experimento *E* y *S* el espacio muestral asociado con el experimento .
Una función *X* que asigna a cada uno de los elementos *s* Є *S* un número real *X (s*) se llama variable aleatoria.

Además existen dos tipos de variables aleatorias que son ;

Discretas; Sea *X* una variable aleatoria, si el número de valores posibles de X (esto es en el recorrido) es finito o infinito numerable la llamamos variable aleatoria discreta.

Continuas ; Se dice que *X* es una variable aleatoria continua si existe una función f llamada función de densidad de probabilidad de x que satisface las siguientes condiciones ;

a) $f_x(X) \geq 0$

b) $\int_{-\infty}^{\infty} Fx^{(x)}\, dx = 1$

c) Para cualquier a, b tal que $-\infty < a < b$= tenemos $P(a \leq X \leq b) = \int_a^b fx^{(x)} dx$

TEORÍA DE CONJUNTOS

Este apartado trata algunas de las ideas y conceptos elementales de la teoría de conjuntos que serán necesarios para una introducción moderna a la teoría de la probabilidad.

CONJUNTOS ELEMENTOS

Se llama conjunto a una lista o colección bien definida de objetos; los objetos comprendidos en un conjunto son llamados elementos o miembros.
Escribimos;

p Є *A* si *p* es un elemento del conjunto *A*

Si cada elemento de a pertenece también a un conjunto B, esto es, si p ЄA implica p Є B, entonces se dice que A es subconjunto de B , o que está contenido en B; esto se denota por:

A C B o B C A

Dos conjuntos son iguales si cada uno está contenido en el otro, esto es;

A = B si y sólo si A C B y B C A.

A menos que otra cosa se establezca, todos los conjuntos en una investigación se suponen subconjuntos de un conjunto fijo llamado *conjunto universal* denotado por "U" . También usamos el símbolo Ø para indicar el conjunto vacío o nulo, esto es, el conjunto que no contiene elementos; este conjunto se considera como un subconjunto de cualquier otro conjunto. Así para cualquier conjunto A, tenemos

Ø C *A* C *U*

CONJUNTOS FINITOS Y CONTABLES

Los conjuntos pueden ser finitos o infinitos. Un conjunto es finito si está vacío o si consta exactamente de *n* elementos en donde n es un entero positivo; de otra manera es infinito.

Un conjunto es contable si es finito o si sus elementos pueden ser ordenados en forma de sucesión, en cuyo caso se dice que es contablemente infinito; de lo contrario el conjunto es no contable.

CONJUNTO PRODUCTO

Sean *A* y *B* dos conjuntos. El conjunto producto de *A* y *B*, expresado por *A* X *B*, está formado por todas las parejas ordenadas (a,b) donde $a \in A$ y $b \in B$

$$A \times B = \{ (a,b) : a \in A,\ b \in B$$

El producto de un conjunto por sí mismo *A* X *A* se denota por A^2

PROBABILIDAD CONDICIONAL

Si E_1 y E_2 son dos sucesos, la probabilidad de que ocurra E_2 se denota por

P{E_2 / E_1 } o P{E_2 dado E_1 } y se llama *probabilidad condicional* de E_2 dado que E_1 se ha presentado.

Si la ocurrencia o no ocurrencia de E_1 no afecta la probabilidad de ocurrencia de E_2 , entonces: P {E_1 / E_2} = { E_2 } y se dice que E_1 y E_2 son *sucesos independientes* ; si no ocurre esto los procesos se dicen *dependientes.*

Si se denota por E_1 y E_2 llamado a veces suceso compuesto se tiene ;

$$P\{E_1E_2\} = P\{E_1\}\, P\{E_2\} \text{ para sucesos independientes}$$

Para tres sucesos E_1,E_2,E_3 se tiene

$$P\{E_1E_2E_3\} = P\{E_2 / E_1\}\ P\{E_3 / E_1 E_2\}$$

Es decir , la probabilidad de ocurrencia de E_1, E_2, E_3 es igual a la probabilidad de E_1 por la probabilidad de ocurra E_2, dado que ha ocurrido E_1, por la probabilidad de que ocurra E_3 dado que ha ocurrido E_1 y E_2 , En particular,

$$P\{E_1E_2E_3\} = P\{E_1\}\ P\{E_2\}\ P\{E_3\} \text{ para sucesos independientes}$$

En general , si $E_1,E_2,E_3.....E_n$ son n sucesos independientes , cuyas probabilidades respectivas son $p_1, p_2,p_3.....p_n$ entonces la probabilidad de ocurrencia de $E_1,E_2,E_3.....E_n$ es $p_1, p_2,p_3.....p_n$

INDEPENCIA

Se dice que un evento *B* es independiente de un evento *A* si la probabilidad de que *B* suceda no está influenciada porqué *A* haya o no sucedido. En otras

palabras, si la probabilidad de *B* iguala la probabilidad condicional de *B* dado *A: P(B) = P(B| A).* Ahora sustituyendo *P(B)* por *P(B | A)* en el teorema de la multiplicación $P(A \cap B) = P(A)\, P(A \mid B)$, obtenemos;

$$P(A \cap B) = P(A)\, P(B)$$

Usamos la definición anterior como nuestra definición formal de independencia

DEFINICIÓN: *A* Y *B* son eventos independientes si $P(A \cap B) = P(A)\, P(B)$; de otro modo son dependientes.

PERMUTACIONES

Una permutación de *n* objetos diferentes tomados de *r* en *r* es una ordenación de *r* objetos entre los *n* dados y atendiendo a la citación de cada objeto en la ordenación. El número de permutaciones de n objetos tomados de *r* en *r* se representa por ${}_nP_r$, *P(n,r*) y viene dado por;

$${}_nP_r = n(n-1)\, n(\, n-2) \ldots (\, n-r+1) = n! / (n-r)\,!$$

COMBINACIONES

Una combinación de *n* objetos diferentes tomados de *r* en *r* es una selección de *r* de los *n* objetos sin atender la ordenación de los mismos. El número de combinaciones de *n* objetos se representan por ${}_nC_r$, *C(n,r),* ó $C^n{}_r$ y viene dad por :

$${}_nC_r = \underline{n(N-1).....(n-r+1)} \qquad n! / r!\,(n - r!)$$

$$r!$$

PRUEBAS REPETIDAS INDEPENDIENTES

Sea *S* un espacio finito de probabilidad, por n pruebas repetidas o independientes, significa que S es espacio de probabilidad T que consta de *n-uplas* o elementos de *S* con la probabilidad de una *n-nupla* definida como el producto de las probabilidades de sus componentes:

$$P((S_1, S_2 \ldots\ldots , S_n) = P(S_1)\, P(S_2)\, \ldots..P(S_n)$$

PARTICIONES Y TEOREMA DE BAYES

Supongamos que los eventos $A_1, A_2 \ldots.. A_n$ forman una partición de espacio muestral *S;* esto es, que los eventos A_i son mutuamente exclusivos y su unión es S. Ahora sea B un evento. Entonces

$$B = S \cap B = (A_1 U A_2 U \ldots\ldots\ldots\ldots .. U A_n) \cap B$$

$$=(A_1 \cap B)\, U\, (A_2 \cap B)\ldots..U(A_n \cap B)$$

Donde las $A_i \cap B$son eventos mutuamente exclusivos. En consecuencia;

$$P(B)= P(A_1 \cap B) + P(A_2 \cap B) + \ldots\ldots..+ P(A_n \cap B)$$

Luego por el teorema de la multiplicación

$$P(B) = P(A_1)\, P(B \mid A_1) + P(A_2)\, P(B \mid A_2) + \ldots\ldots\ldots..+P(A_n)\, P(B \mid A_n)$$

Por otra parte, para cualquier *I* , la probabilidad condicional de A_i dado B se define como :

$$P(A_1 \mid B) = P(A_1 \cap B) / P(B)$$

En esta ecuación usamos *i* para reemplazar, $P(B)$ y usamos $P(A_1 \cap B) = P(A_i)\ P(B \mid A_i)$ para reemplazar $P(A_i \cap B)$, obteniendo así el Teorema de bayes

TEOREMA DE BAYES

Sopóngase que $A_1, A_2, \ldots\ldots A_n$ es una partición de S y B es cualquier evento. Entonces para cualquier *i.*

$$P(A_1 \mid B) = P(A_i)\ P(B \mid A_i) / P(A_i)\ P(B \mid A_i) + P(A_2)\ P(B \mid A_2) + \ldots\ldots + P(A_n)\ P(B \mid A_n)$$

DISTRIBUCIONES DE PROBABILIDAD DISCRETA

Si una variable X puede tomar una serie de valores discretos $X_1, X_2, \ldots\ldots X_k$, con probabilidades respectivas, $p_1, p_2, \ldots\ldots p_k$, donde la sumatoria de todas estas probabilidades es igual a 1, se dice que ha sido definida una X para una probabilidad discreta. La función $p(X)$ que toma los valores respectivos $p_1, p_2, \ldots\ldots p_k$, para $X = X_1, X_2, \ldots\ldots X_k$, se llama función de probabilidad o función de frecuencia *X.* Como X puede tomar ciertos valores con probabilidades dadas, se llama a veces variable aleatoria discreta. Una variable aleatoria se conoce también como variable de probabilidad o variable estocástica.

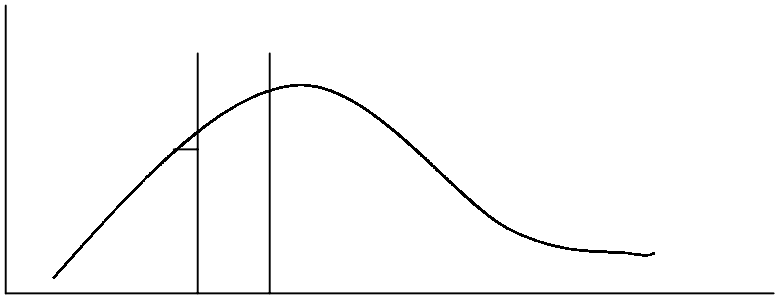
DISTRIBUCIONES DE PROBABILIDAD CONTINUA

Las ideas anteriores igualmente pueden entenderse al caso que la variable X puede tomar una serie de valores continuos. El polígono de frecuencias de una muestra llega a ser, en el caso límite de una población una curva continua, tal

como la que se muestra en la figura, cuya ecuación es *Y = p(X).* El área total bajo esta curva limitada por el eje *X* es igual a uno, y el área bajo la curva y entre las rectas *X = a y X =b* (área sombreada de la figura) da la probabilidad de que *X* se encuentre entre a y b lo que se puede representar por *P{a < X < X).*

Se conoce como *p(X)* como una función de la probabilidad, o brevemente como función de densidad, y cuando tal función es dada se dice que la distribución de probabilidad continua para *X* ha sido definida. La variable *X* se llama también variable aleatoria continua.

Como en el caso discreto, se definen las distribuciones de probabilidad acumulada y las funciones de distribución asociadas a ellas.



DISTRIBUCIONES ESPECIALES

DISTRIBUCIÓN BINOMIAL

Si *P* es la probabilidad de un suceso en un solo ensayo y *p = 1 = p* es la probabilidad de que el suceso no ocurra en un solo ensayo entonces la probabilidad de que el suceso se presente exactamente en n ensayos está dada por:

$$P[X = x](n, x)\, p^{x}\, q^{n-x} = n!/x!(n-x)\, p^{x}\, q^{n-x} \quad x = 0,1,2,3.....,n$$

En donde $n \in N$ y $0 \leq P \leq 1$, $X \sim \beta(n,P)$

PROPIEDADES DE LA DISTRIBUCIÓN BINOMIAL

A).-Note que para cada *n* y *P* se tiene una distribución. Se dice que *n* y *P* son los parámetros de la distribución binomial

B).- La media de la binomial es $\mu = nP$

C).- La varianza de la binomial es $\sigma^2 = npq$

D),- La desviación típica de la binomial es $\sigma = \sqrt{npq}$

DISTRIBUCIÓN HIPERGEOMÉTRICA

Supóngase que tenemos un lote de $N = m + n$ artículos de los cuales m tiene una característica, y $n = N-m$ no tienen la característica.

Supóngase que escogemos al azar r artículos de ese lote $\{¨(r \leq N), (r \leq n), (r \leq m)\}$ sin sustitución.

Sea *X* el número de artículos con la característica puesto que $X = X$ sí y solo obtenemos *x* artículos con la característica y exactamente $(r - x)$ artículos sin la característica. Entonces;

$$P[X = x] = \binom{m}{x}\binom{n}{r-x} / \binom{m+n}{r}$$

$$X = 0,1,2,3,\ldots\ldots\ldots n$$

D.O.M.

En donde $r \leq m + n$, $r \leq n$, $r \leq m$, con $r, n, m \in N$

DISTRIBUCIÓN POISSON

Sea X una variable aleatoria que los valores posibles 1,2,3,4, ...n sí

$$P(X = x) = e^{-\lambda} \lambda^x / x!$$

$$x = 0,1,2,3.....,n$$

decimos que x tiene una distribución *Poisson* con parámetro $\lambda > 0$

TEOREMA;

Si x tiene una distribución de *Poisson* con parámetro λ entonces;

$$E(X) = \lambda \text{ y } Var(X) = \lambda$$

DEMOSTRACION;

TEO; Sea x una variable aleatoria distribuida binomialmente con parámetro *P*(con base en n repeticiones del experimento). Esto es;

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Supóngase que cuando n $\longrightarrow \infty$ $np = \lambda$ *o* equivalente cuando n $\longrightarrow \infty$ $p \longrightarrow$ 1 tal que

$Np \longrightarrow \lambda$ bajo estas condiciones tenemos;

$$\lim_{n \to \infty} P(X = x) = e^{-\lambda} \lambda^x / x!$$

CARACTERÍSTICAS DE LA DISTRIBUCIÓN POISSON;

A) la media para la Distribución POISSON es igual a np, es decir m ($\mu = np$)

B) de la misma manera la varianza para esta distribución es np o sea = m (σ^2= *np)*

C) por ultimo la desviación estándar es igual a σ =√*np*

EJEMPLOS PACTICOS

PERMUTACIONES

¿De cuántas formas es posible permutar 7 vacas en corraletas individuales de una nave con capacidad para 20?

A) Si solo se ocupan 12 jaulas;

$$P_x^n = \frac{n}{(n-x)!} =$$

$$P_7^{12} = \frac{12!}{(12-7)!} = 3{,}991{,}680$$

Se pueden acomodar de 3,991,680 formas.

B) si se ocupan las jaulas con número non;

$$P_7^{10} = \frac{10!}{(10-7)!} = 609{,}800$$

Por lo visto se pueden acomodar de 609,800 maneras.

C) Si se ocupan todas las jaulas;

$$P_7^{20} = \frac{20}{(20-7)} = 390{,}200{,}800$$

O sea se pueden permutar de 390,200,300 formas.

COMBINACIONES

En un termo de inseminación se tienen pajillas de semen dispuestas de la siguiente manera:

CANTIDAD	TORO
12	UA-903
5	UA-904
4	UA-108
2	UA-101

Por otra parte se desean 9 pajillas de las cuales deben ser;

CANTIDAD	TORO
2	UA-903
2	UA-904
3	UA-108
2	UA-101

A) ¿De cuántas formas se pueden tomar?

$$C_2^{12} C_2^5 C_3^9 C_2^7 = 1{,}164{,}240$$

Esto quiere decir que se pueden acomodar de 1,164,240 formas.

B) ¿De cuántas formas es posible tomarlas si 5 del TORO UA-903 no pueden ser incluidos en el arreglo

$$C_2^7 C_2^5 C_3^9 C_2^7 = 3{,}700{,}440$$

O sea que se pueden arreglar de 3,700,440

C) ¿De cuántas formas es posible tomarlas si 4 del TORO UA-101 no pueden estar en el arreglo

$$C_2^{12} C_2^5 C_3^9 C_2^3 = 166{,}320$$

Es decir que se pueden tomar de 166,320 formas

PROBABILIDAD

En un invernadero se encuentran plantas de durazno dispuestas de la siguiente manera;

PERIODO	VARIEDAD 1	VARIEDAD 2	CRIOLLO
TARDÍO	15	8	12
TEMPRANO	7	10	14
	22	**18**	**26**
			66

Calcular las siguientes probabilidades;

A).- Que al tomar un aplanta la azar, esta sea tardío o variedad dos:

35/66 + 18/66 – 8/66 = .6818 ó **68.18%**

B).- Que al tomar una planta al azar sea variedad 1

26/66 +22/6 = .7272 ó **72.72%**

C).- Que al tomar tres plantas al azar sea criollo

(26/66) (25/65) (24/64) = 15600/274550 .0568 ó 5.68%

D).- Que al tomar 2 plantas al azar y con reemplazo las dos sean tardío

3(35/66) (35/66) = .28122 ó 28.12%

E).- Que al tomar 3 plantas en orden sean V1, V2, V3

(22/66) (18/65) (26/64) = .036931818 ó 36.93%

PRUEBAS REPETIDAS INDEPENDIENTES

La probabilidad de Neumonía en terneros Holstein en un establo es de 8%, si se toma al azar una muestra de cinco animales A).- ¿cuál es la probabilidad de que dos contengan Neumonía, B).- A lo más uno contenga la enfermedad, C).- Al menos cuatro no tengan Neumonía.

A).- Probabilidad de que dos contengan Neumonía

p = .08 n = 5

q = .92 x = 2

P(que 2 contengan neumonía) = C^5_2 $.08^2$ $.92^{5-2}$ =

(10) (.0069) (.778688) = .**0498**

Esto quiere decir que la probabilidad de 2 animales tengan Neumonía es de 4.98%

B).- Probabilidad de que a lo mas 1 tenga Neumonía

P(de que al menos uno tenga Neumonía) = (C^5_0) $(.92^0)$ $(.92^{5-0})$ + (C^5_1) $(.08^1)$ $(.92^{5-1})$ = .28652

(1) (0) (.6590) + (5) (.08) (.7163) = .**28652**

Esto se traduce en la probabilidad de a lo más un animal contenga Neumonía es de 28.65 %

C).- Probabilidad de que a lo menos 4 no tengan Neumonía

P(de que al menos cuatro no tengan Neumonía) =

C^5_4 $(.92^4)$ $(.08^{5-4})$ + C^5_5 $(.92^5)$ (08^{5-5}) =

5(.7163) (.08) + 1(.6590) (0) = .**28652**

Esto significa que la probabilidad de que al menos cuatro terneros no contengan Neumonía es de 28.65 %.

TEOREMA DE BAYES

La compra de mamilas para el sistema de ordeña de un establo se efectúa de la siguiente manera:

FABRICANTE	VOLUMEN	DEFECTUOSAS
CORMORAN	40	7

ALBA	35	4
MEX-SOL	35 110	9 20

A).- Si al azar se toma una pieza ¿cuál es la probabilidad de que sea defectuosa y de Mex-sol

Cormoran	Mexsol	Alba
7/20	4/20	9/20
.40	.35	.35

$$P(\text{defectuosas/ Mexsol}) = \frac{(9/20)\,(.35)}{(7/20).40 + (4/20).35 + (9/20).35}$$

P = .4285 ó 42.85%

Es decir que la probabilidad de que la mamila sea defectuosa y además fabricada por Mexsol es de 42.85 %

DISTRIBUCIÓN HIPERGEOMÉTRICA

Siete de cada 50 toretes Holstein de selección no son de primera, si al azar se toma una muestra de 4 ¿Cuál es la probabilidad de:

A) Que dos no sean de primera

$$P = \frac{C_2^7 C_2^{43}}{C_4^{50}}$$ **= 0.082 ó 8.2 % de probabilidad de que 2 no sean de primera**

B) Que al menos dos sean primera

$$P = \frac{C_3^{43} C_1^7}{C_4^{50}} + \frac{C_4^{43} C_0^7}{C_4^{50}} = 0.545$$

O sea el 54.5% de probabilidad de que al menos dos terneros sean de primera.

Si 8 de cada 89 muestras de forraje colectado en el estado de Coahuila presentan deficiencia de fósforo y en forma azarizada se toma una muestra de tamaño 6

¿ Cuál es la probabilidad de que?

A) Dos presenten deficiencia de fósforo

$$P(n_1) = \frac{C_{n-n-1}^{N-N-1} C_{n1}^{N1}}{C_n^N} =$$

P(2 deficiente)= $\frac{C_2^8 C_4^{81}}{C_6^{89}} = \frac{1663740}{581106988} =$ ***.081***

Esto quiere decir que la probabilidad de que dos muestras de forraje tengan deficiencia de fósforo es de 8.10%

C) Que al menos una presente deficiencia de fósforo;

P = (al menos una deficiente)

$$P\ \frac{C_5^{81}C_1^8}{C_6^{89}}+\frac{C_4^{81}C_2^8}{C_6^{89}}+\frac{C_3^{81}C_3^8}{C_6^{89}}+\frac{C_2^{81}C_4^8}{C_6^{89}}+\frac{C_1^{81}C_5^8}{C_6^{89}}+\frac{C_0^{81}C_6^8}{C_6^{89}} = .441397$$

Es decir que la probabilidad de al menos 1 presente deficiencia de fósforo es de 44.13%

DISTRIBUCIÓN POISSON

La probabilidad de fiebre de embarque en ganado bovino con vacuna es de 1.2%, por otra parte se estima una muestra al azar de 42 animales ¿Cuál es la probabilidad de que 2 tengan fiebre de embarque y qué a lo más uno presente la enfermedad;

A) Probabilidad de que dos tenga fiebre de embarque;

p = 0.012

n = 42

m = np = (0.012)(42)= .504

$$F(x) = \frac{m^{x} e^{m}}{X!}$$

$$P(2) = \frac{.0504^{2} e^{-0.504}}{2!}$$

*P(2)***= *0.076 ö 7.6%***

Esto quiere decir que la probabilidad de 2 animales padezcan la enfermedad es de 7.67%.

D) Probabilidad de que a lo mas 1 tenga fiebre de embarque

P = 0.012
n = 42
m = np = (0.012)(42)= .504

$$F(0,1) = \frac{0.504^{0} e^{-0.504}}{01} + \frac{0.504^{1} e^{-0.504}}{1!} =$$

$$F(0,1) = .604 + .304 = \quad \textbf{.9058}$$

Es decir que existe un 90.85% de probabilidad de que se presente la fiebre de embarque en por lo menos 1 animal.

La probabilidad de preñez en el primer celo posparto en bovinos es muy pequeña1.7%. Si al azar se toma una muestra de 42 animales, con estas características defina la probabilidad de que;

A) 2 queden preñadas

$$F(X) = \frac{m^{x} e^{-m}}{X!} =$$

p =.017

x = 2

n= 42

m = np = .714

$$F(2) = \frac{.714^{2} e^{-0.714}}{2!} = .1248$$

O sea que la probabilidad de que dos hembras queden preñadas es de 12.48%

B) A lo menos una quede preñada

P=.017

X= 2

n= 42

m= np= .714

$$F(0,1) = \frac{0.714^{0} e^{-0.714}}{0!} + \frac{0.714^{1} e^{-0.714}}{1!} =$$

.4896 + 34.96 = .8392

Esto nos indica que la probabilidad de que lo más una quede preñada es de 83.92%

DISTRIBUCIÓN BINOMIAL

Si la probabilidad de Brucelosis en cabras en una zona templada es de 8.3%. Si al azar se toma una muestra de 8 animales defina la probabilidad de que;

A) Dos no tengan brucelosis;

$$F(X) = C_x^n p^x q^{n-x}$$

p = .083 n = 8

q = .917 x = 2

$P(2) = C_2^8 (.917)^2 (.083)^{8-2}$ = 28(.8400889)(.000000326)= .000007679

Como se observa es muy baja la probabilidad

B) A lo más dos tengan Brucelosis;

$$C_0^8 (.083)^0 (.917)^{8-0} + C_1^8 (.083)^1 (.917)^{8-1} + C_2^8 (.083)^2 (.917)^{8-2} =$$

1(1)(4999)+8(.083)(.5452)+28(.006889)(.5946)= .9765

Esto quiere decir que la probabilidad de que a lo mas dos tengan Brucelosis es muy alta ya que es del 97.65%

C) Tres padezcan de Brucelosis

$$P(3) = C_3^8 (.083)^3 (.917)^{8-3}$$

(56) (.000571) (.6484) = .0207

La probabilidad de tres animales con Brucelosis es de 2.07 %

LITERATURA REVISADA

Bioestadística: Métodos y Aplicaciones. Universidad de Málaga. http://www.bioestadistica.uma.es/libro/

Castillo P.J, J.G.Arias. 1998. Estadística inferencial básica. Grupo editorial Ibero América. México.

Cochran. G. William. 1980. Diseños experimentales. Editorial Trillas. México

Infante G. S. 1997. Métodos estadísticos. Editorial Trillas. México.

Kreyszig Erwin. 1979. Introducción a la estadística matemática. Editorial Limusa. México

Montgomery D. C.1991.Diseño y análisis de experimentos. Grupo editorial Iberoamérica. México.

Ostle, B. 1965. Estadística aplicada. Primera edición. Editorial Limusa. México.

Rodríguez del A. J. 1991. Métodos de investigación pecuaria. Editorial Trillas. México.

Snedecor W. George, W. G. Cochran. 1979 métodos estadísticos. Editorial Continental. México.

Steel G.D Robert, J. H. Torrie.1981. 2ª. Principles and procedures of statistics a biometrical approach. 2ª. Ed. Editorial Mc Graw-Hill. USA.

Walpole, R. E.1992. Probabilidad y estadística. Cuarta Edición. Editorial Mc GrawHill. México.

DISTRIBUCIÓN NORMAL O GAUSSIANA

INTRODUCCION

Para hacer una definición rigurosa de la probabilidad, necesitamos precisar ciertas leyes o axiomas que deba cumplir una función de probabilidad. Intuitivamente estos axiomas deberían implicar, entre otras, las siguientes cuestiones, que nos parecen lógicas en términos de lo que se puede esperar de una función de probabilidad:

- La probabilidad sólo puede tomar valores comprendidos entre 0 y 1(no puede haber sucesos cuya probabilidad de ocurrir sea del 200% ni del –5.
- La probabilidad del suceso seguro es 1, es decir, el 100%.
- La probabilidad del suceso imposible debe ser 0.
- La probabilidad de la intersección de dos sucesos debe ser menor o igual que la probabilidad de cada uno de los sucesos por separado, es decir:

$$\mathcal{P}_{rob}[A \cap B] \leq \mathcal{P}_{rob}[A]$$

$$\mathcal{P}_{rob}[A \cap B] \leq \mathcal{P}_{rob}[B]$$

- La probabilidad de la unión de sucesos debe ser mayor que la de cada uno de los sucesos por separado:

$$\mathcal{P}_{rob}[A \cup B] \geq \mathcal{P}_{rob}[A]$$

$$\mathcal{P}_{rob}[A \cup B] \geq \mathcal{P}_{rob}[B]$$

  Más aún, si los sucesos son disjuntos (incompatibles) debe ocurrir que

$$A \cap B = \emptyset \quad \Longrightarrow \quad \mathcal{P}_{rob}[A \cup B] = \mathcal{P}_{rob}[A] + \mathcal{P}_{rob}[B]$$

- La probabilidad del suceso contrario de *A*, debe valer

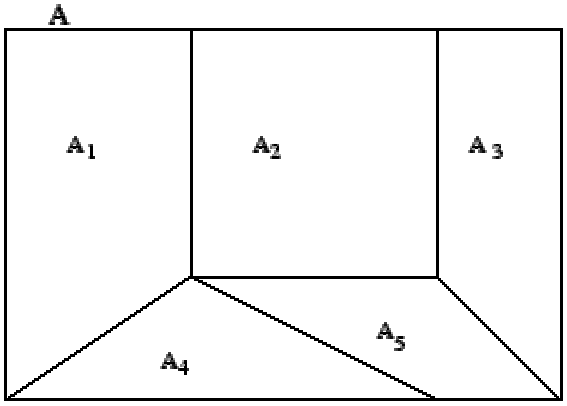
Prob[A] = 1 - Prob[A]. Esto en realidad puede deducirse del siguiente razonamiento:

$$A\cap\overline{A}=\emptyset \;\Rightarrow\; 1=\mathcal{P}_{rob}[E]=\mathcal{P}_{rob}[A\cup\overline{A}]=\mathcal{P}_{rob}[A]+\mathcal{P}_{rob}[\overline{A}] \;\Rightarrow\; \mathcal{P}_{rob}[\overline{A}]=1-\mathcal{P}_{rob}[A]$$

La probabilidad de la unión numerable de sucesos disjuntos es la suma de sus probabilidades (figura 1).

$$A_1, A_2, \ldots, A_n, \ldots \in \mathcal{A} \Longrightarrow \mathcal{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathcal{P}[A_i]$$

Figura:1 Si $A = A_1 \cup A_2 \cup \cdots$ con $A_i \cap A_j = \emptyset$, entonces $\mathcal{P}[A] = \mathcal{P}[A_1] + \mathcal{P}[A_2] + \cdots$



siguiendo esos puntos:

La función de probabilidad debe calcularse sobre subconjuntos de *E*. No es estrictamente necesario que sean *todos*, pero si es necesario que si se puede calcular sobre un conjunto, lo pueda ser también sobre su complementario, y que si se puede calcular sobre dos conjuntos *A* y *B*, que también se pueda calcular sobre su unión y su intersección. Para ello introduciremos el concepto de σ*-álgebra de sucesos*, que será una clase de subconjuntos de *E* sobre los que podamos aplicar las reglas de la probabilidad.

Entre las leyes que debe cumplir una función de probabilidad y que se han escrito antes, se ha observado que algunas son redundantes, ya que se pueden deducir de las demás. Con la definición axiomática de la probabilidad se pretende dar el menor conjunto posible de estas reglas, para que las demás se deduzcan como una simple consecuencia de ellas.

Se precisa entonces los conceptos de *σ-álgebra de sucesos* y de *probabilidad*.

CONCEPTO DE σ-ÁLGEBRA DE SUCESOS

Sea $\mathcal{A}$una clase no vacía formada por ciertos subconjuntos del espacio muestral *E*. Diremos que esta clase es un σ-álgebra de sucesos si los sucesos complementarios de aquellos que están en $\mathcal{A}$también están en $\mathcal{A}$, así como sus uniones numerables (sean finitas o infinitas). Esto se puede enunciar como:

$$\mathcal{A} \text{ es un } \sigma\text{ –álgebra} \iff \begin{cases} \forall A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A} \\ \forall A_1, \ldots, A_n \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{A} \end{cases}$$

La introducción de la definición de *σ-álgebra* puede parecer innecesaria a primera vista, ya que es una clase formada por subconjuntos de *E* que verifican ciertas propiedades relativas a la complementariedad y a las uniones finitas que ya verifica de antemano el conjunto denominado partes de *E*, *P*(*E*), formado por todos los subconjuntos de *E*. Cuando el conjunto *E* de los posibles resultados de un experimento aleatorio sea finito, normalmente consideraremos como σ-álgebra de sucesos al conjunto *P*(*E*). Esto ocurre cuando por ejemplo realizamos el experimento aleatorio de lanzar un dado:

$$E = \{1,2,3,4,5,6\}$$

$$\mathcal{A} = P(E) = \{\emptyset, E, \{1\}, \{2\}, \ldots, \{1,2\}, \{1,3\}, \ldots, \{1,2,3\}, \ldots\}$$

Cuando *E* es infinito no numerable, la estructura del conjunto *P*(*E*) puede presentar propiedades extremadamente engorrosas. Entonces es más

conveniente utilizar como σ-álgebra un subconjunto más pequeño suyo, pero no tanto que no nos permita realizar las operaciones de complementariedad o de uniones finitas que se precisan en la definición de un σ-álgebra. Por ejemplo, si realizamos el experimento aleatorio de esperar el tiempo que hace falta para que un átomo de carbono catorce, C^{14}, se desintegre de modo natural, se tiene que

$$E = \mathbb{R}^+,$$

Sin embargo, el σ-álgebra de sucesos que consideramos no es $P(\mathbb{R}^+)$, que es una clase demasiado compleja para definir sobre sus elementos una medida de probabilidad. En su lugar consideramos el σ-álgebra formada por todos los intervalos, abiertos o cerrados, y sus uniones finitas, lo que, por

$$\mathcal{A} = \{\emptyset, \mathbb{R}^+, (2,3), (4,5] \cup [8,+\infty), \ldots\}$$

supuesto incluye a los puntos de $\mathbb{R}^+$, ya que por ejemplo

$$\{2\} = [2,2].$$

Este tipo de conjuntos (los intervalos) son los que nos interesan en la práctica (Walpole, 1992).

PROBABILIDAD CONDICIONADA E INDEPENDENCIA DE SUCESOS

Sea *B* ⊂ *E* un suceso aleatorio de probabilidad no nula, P[B]>0. Para cualquier otro suceso A ⊂ *E*, llamamos probabilidad condicionada de *A* a *B* a la cantidad que representamos mediante $P[A_{|B}]$ o bien P_B [A] y que se calcula como:

$$\boxed{\mathcal{P}[A_{|B}] = \frac{\mathcal{P}[A \cap B]}{\mathcal{P}[B]}}$$

Ejemplo

Se lanza un dado al aire ¿Cuál es la probabilidad de que salga el número 4? Si sabemos que el resultado ha sido un número par, ¿se ha modificado esta probabilidad?

Solución:

El espacio muestral que corresponde a este experimento es:

$$E = \{1, 2, 3, 4, 5, 6\}$$

y se ha de calcular la probabilidad del suceso $A = \{4\}$. Si el dado no está trucado, todos los números tienen la misma probabilidad de salir, y siguiendo la definición de probabilidad de Laplace,

$$\begin{aligned} \mathcal{P}[A] &= \frac{\text{casos favorables}}{\text{casos posibles}} \\ &= \frac{\text{número de elementos en } \{4\}}{\text{número de elementos en } \{1, 2, 3, 4, 5, 6\}} \\ &= \frac{1}{6} \end{aligned}$$

Obsérvese que para calcular la probabilidad de *A* según la definición de Laplace se ha tenido que suponer previamente que todos los elementos del espacio muestral tienen la misma probabilidad de salir, es decir:

$$\mathcal{P}[1] = \mathcal{P}[2] = \mathcal{P}[3] = \mathcal{P}[4] = \mathcal{P}[5] = \mathcal{P}[6]$$

Por otro lado, si ha salido un número par, de nuevo por la definición de probabilidad de Laplace tendríamos

$$\begin{aligned}\mathcal{P}_{\mathbf{par}}[4] &= \frac{\textbf{casos favorables}}{\textbf{casos posibles}} \\ &= \frac{\textbf{número de elementos en } \{4\}}{\textbf{número de elementos en } \{2,4,6\}} \\ &= \frac{1}{3}\end{aligned}$$

Esta misma probabilidad se podría haber calculado siguiendo la definición de la probabilidad condicionada, ya que si escribimos

$$\begin{aligned} A = \{4\} \quad &\Rightarrow \quad \mathcal{P}[A] = \frac{1}{6} \\ B = \{2,4,6\} \quad &\Rightarrow \quad \mathcal{P}[B] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \\ A \cap B = \{4\} \quad &\Rightarrow \quad \mathcal{P}[A \cap B] = \frac{1}{6} \end{aligned}$$

y entonces

$$\mathcal{P}_{\mathbf{par}}[4] = \mathcal{P}_B[A] = \mathcal{P}[A_{|B}] = \frac{\mathcal{P}[A \cap B]}{\mathcal{P}[B]} = \frac{1/6}{1/2} = \frac{1}{3}$$

que por supuesto coincide con el mismo valor que calculamos usando la definición de probabilidad de Laplace. Obsérvese que según la definición de probabilidad condicionada, se puede escribir la probabilidad de la intersección de dos sucesos de probabilidad no nula como

$$\mathcal{P}[A \cap B] = \begin{cases} \mathcal{P}[A] \cdot \mathcal{P}[B_{|A}] \\ \mathcal{P}[B] \cdot \mathcal{P}[A_{|B}] \end{cases}$$

O sea, la probabilidad de la intersección de dos sucesos, es la probabilidad de uno cualquiera de ellos, multiplicada por la probabilidad del segundo *sabiendo que* ha ocurrido el primero. Si entre dos sucesos no existe ninguna relación cabe esperar

que la expresión ``*sabiendo que*'' no aporte ninguna información. De este modo introducimos el concepto de independencia de dos sucesos *A* y *B* como:

$$A \text{ es independiente de B} \iff \mathcal{P}[A \cap B] = \mathcal{P}[A] \cdot \mathcal{P}[B]$$

Esta relación puede ser escrita de modo equivalente, cuando dos sucesos son de probabilidad no nula como

$$A \text{ es independiente de B}\ , \mathcal{P}[A] \neq 0 \neq \mathcal{P}[B] \quad \Leftrightarrow \quad \begin{cases} \mathcal{P}[A] = \mathcal{P}[A_{|B}] \\ \text{o equivalentemente} \\ \mathcal{P}[B] = \mathcal{P}[B_{|A}] \end{cases}$$

DISTRIBUCIÓN NORMAL O GAUSSIANA

La *distribución gaussiana*, recibe también el nombre de *distribución normal*, ya que una gran mayoría de las v.a continuas de la naturaleza siguen esta distribución. Se dice que una v.a. X sigue una distribución normal de parámetros μ y σ^2 lo que representamos del modo

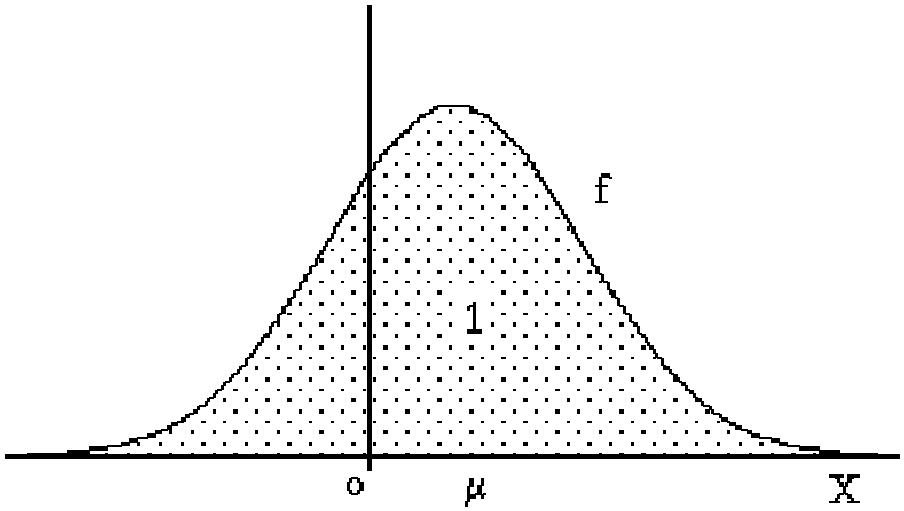
$$X \rightsquigarrow \mathbf{N}\left(\mu, \sigma^2\right)$$

si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad \forall x \in \mathbb{R}$$

La forma de la función de densidad es la llamada *campana de Gauss*.

Figura: 2. Campana de Gauss o función de densidad de una v.a. de distribución normal. El área contenida entre la gráfica y el eje de abcisas vale 1.



Es un ejercicio interesante comprobar que ésta alcanza un único máximo *(moda)* en μ, que es simétrica con respecto al mismo, y por tanto

$$\mathcal{P}[X \leq \mu] = \mathcal{P}[X \geq \mu] = 1/2$$

,

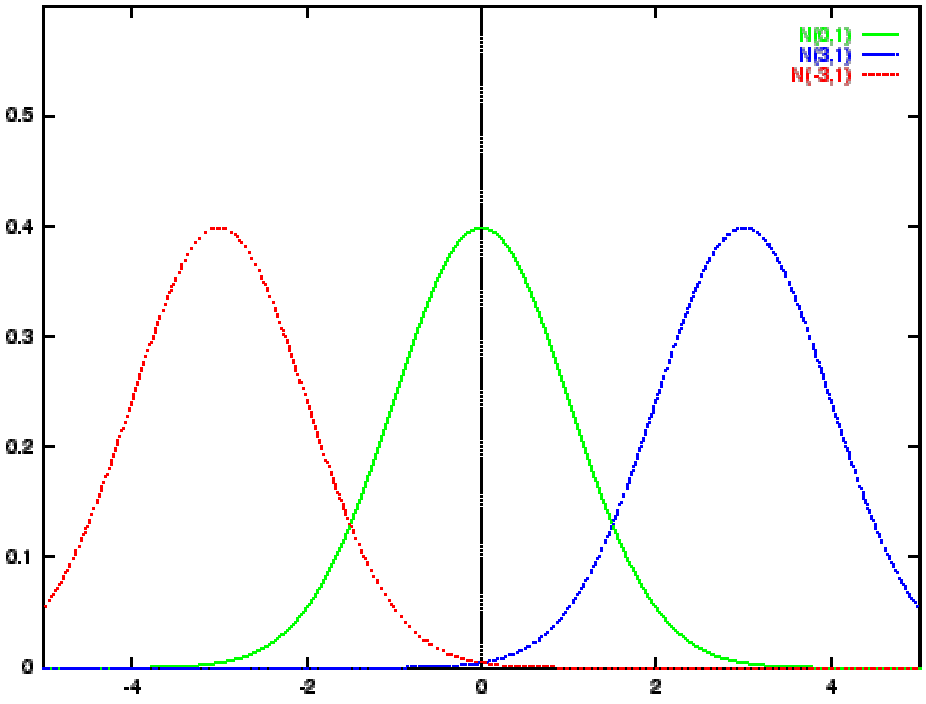
con lo cual en μ coinciden la media, la mediana y la moda, y por último, también interesante, calcular sus puntos de inflexión**.** (Steel, 1986)**.**

El soporte de la distribución es todo $\mathbb{R}$, de modo que la mayor parte de la *masa de probabilidad* (área comprendida entre la curva y el eje de abcisas) se encuentra concentrado alrededor de la media, y las ramas de la curva se extienden asintóticamente a los ejes, de modo que cualquier valor ``muy alejado'' de la media es posible (aunque poco probable).

La forma de la campana de Gauss depende de los parámetros μ y σ.

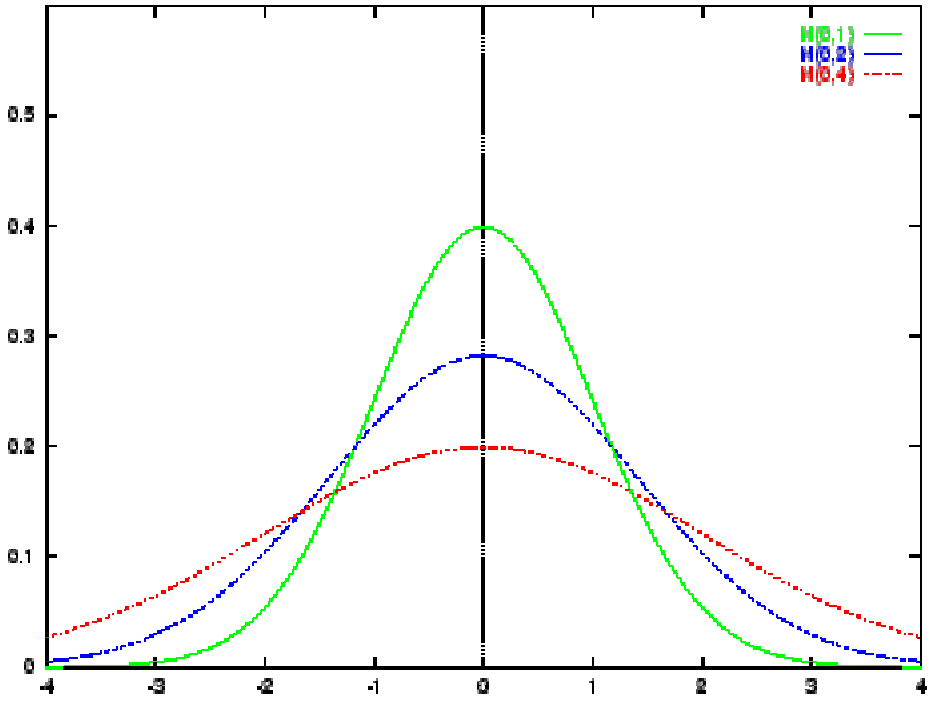
- μ, indica la posición de la campana *(parámetro de centralización)*;

Figura:3 Distribuciones gaussianas con diferentes medias e igual dispersión.



- σ^2 (o equivalentemente, σ) será el parámetro de dispersión. Cuanto menor sea, mayor cantidad de masa de probabilidad habrá concentrada alrededor de la media (grafo de *f* muy apuntado cerca de μ y cuanto mayor sea ``más aplastado" será.

Figura:4. Distribuciones gaussianas con igual media pero varianza diferente.



La función característica de la distribución normal, se comprueba más adelante que es:

$$\phi_X(t) = e^{it\mu - \frac{1}{2}t^2\sigma^2}$$

Como consecuencia, la distribución normal es reproductiva con respecto a los parámetros μ y σ^2, ya que

$$\begin{cases} X \rightsquigarrow \mathbf{N}(\mu_1, \sigma_1^2) \\ \textbf{independientes} \\ Y \rightsquigarrow \mathbf{N}(\mu_2, \sigma_2^2) \end{cases} \Longrightarrow \begin{cases} \phi_X(t) = e^{it\mu_1 - \frac{1}{2}t^2\sigma_1^2} \\ \textbf{independientes} \\ \phi_Y(t) = e^{it\mu_2 - \frac{1}{2}t^2\sigma_2^2} \end{cases}$$

$$\Longrightarrow \quad \phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t) = e^{it(\mu_1+\mu_2) - \frac{1}{2}t^2(\sigma_1^2+\sigma_2^2)}$$

$$\Longleftrightarrow \quad X + Y \rightsquigarrow \mathbf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

A pesar de la utilidad de la ley gausiana, hay que apuntar un hecho *negativo* para esta ley de probabilidad:

La función e^{-x^2} no posee primitiva conocida. Las consecuencias desde el punto de vista práctico son importantes, ya que eso impide el que podamos escribir de modo sencillo la función de distribución de la normal, y nos tenemos que limitar a decir que:

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\,dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

sin poder hacer uso de ninguna expresión que la simplifique. Afortunadamente esto no impide que para un valor de *x* fijo, *F*(*x*) pueda ser calculado. De hecho puede ser calculado con tanta precisión (decimales) como se quiera, pero para esto se necesita usar técnicas de cálculo numérico y ordenadores. Para la utilización en problemas prácticos de la función de distribución *F*, existen ciertas tablas donde se ofrecen (con varios decimales de precisión) los valores *F*(*x*) para una serie limitada de valores x_i dados. Normalmente *F* se encuentra tabulada para una distribución *Z*, normal de media 0 y varianza 1 que se denomina *distribución normal tipificada*:

$$Z \rightsquigarrow \mathbf{N}(0,1) \qquad \Longleftrightarrow f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \forall z \in \mathbb{R}$$

Proposición

Sea $X \rightsquigarrow \mathbf{N}(\mu, \sigma)$. Entonces

$$\begin{aligned} \mathbf{E}[X] &= \mu \\ \mathbf{Var}[X] &= \sigma^2 \\ \phi_X(t) &= e^{it\mu - \frac{1}{2}t^2\sigma^2} \end{aligned}$$

Demostración

Por ser la normal una ley de probabilidad se tiene que

$$\int_{-\infty}^{+\infty} \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

es decir, esa integral es constante. Con lo cual, derivando la expresión anterior con respecto a μ se obtiene el valor 0:

$$\begin{aligned} 0 &= \int_{-\infty}^{+\infty} \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tfrac{1}{\sigma} 2\left(\tfrac{x-\mu}{\sigma}\right) dx \\ &= \frac{1}{\sigma^2}\left[\underbrace{\int_{-\infty}^{+\infty} \tfrac{1}{\sigma\sqrt{2\pi}} x \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx}_{=\mathbf{E}[X]} - \mu \cdot \underbrace{\int_{-\infty}^{+\infty} \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx}_{=1}\right] \end{aligned}$$

luego $\mathbf{E}[X] - \mu = 0$.

Para demostrar la igualdad entre la **Var [X]** y σ^2, basta con aplicar la misma técnica, pero esta vez derivando con respecto a σ^2:

$$0 = -\frac{1}{2}\left[\frac{1}{\sigma^2}\underbrace{\int_{-\infty}^{+\infty} \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx}_{=1} - \frac{1}{\sigma^4}\underbrace{\int_{-\infty}^{+\infty} \tfrac{1}{\sigma\sqrt{2\pi}} (x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx}_{=\mathbf{E}[(X-\mu)^2]=\mathbf{Var}[X]}\right]$$

Luego

$$\frac{1}{\sigma^2} - \frac{1}{\sigma^4}\mathbf{Var}[X] = 0 \Longrightarrow \mathbf{Var}[X] = \sigma^2$$

Para demostrar el resultado relativo a la función característica, consideramos en primer lugar la v.a. tipificada de *X*,

$$Z = \frac{X-\mu}{\sigma} \rightsquigarrow \mathbf{N}(0,1)$$

y calculamos

$$\phi_Z(t) = \int_{-\infty}^{+\infty} e^{itz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(z-it)^2}\, dz}_{\sqrt{2\pi}} = e^{-\frac{t^2}{2}}$$

Como $X = \mu + \sigma U$, por la proposición 5 deducimos que

$$\phi_X(t) = e^{it\mu}\phi_Z(\sigma t) = e^{it\mu - \frac{1}{2}t^2\sigma^2}$$

Aproximación a la normal de la ley binomial

Se puede demostrar **(teorema central del límite)** que una v.a. discreta con distribución binomial, X∿ B(n,p)se puede aproximar mediante una distribución normal si *n* es suficientemente grande y *p* no está ni muy próximo a 0 ni a 1. Como el valor esperado y la varianza de *X* son respectivamente np y npq, la aproximación consiste en decir que

$$X \overset{\approx}{\rightsquigarrow} \mathbf{N}(np, npq)$$

.

El convenio que se suele utilizar para poder realizar esta aproximación es:

$$X \rightsquigarrow \mathbf{B}(n,p) \quad \text{donde} \quad \begin{cases} n > 30 \\ np > 4 \\ nq > 4 \end{cases} \implies X \overset{\approx}{\rightsquigarrow} \mathbf{N}(np, npq)$$

aunque en realidad esta no da resultados muy precisos a menos que realmente *n* sea un valor muy grande o p ≈ p ≈ ½. Como ilustración obsérvense las figuras 5 Y 6.

Figura 5: Comparación entre la función de densidad de una v.a. continua con distribución N(np, npq) y el diagrama de barras de una v.a. discreta de distribución B(n,p) para casos en que la aproximación normal de la binomial es válida. Es peor esta aproximación cuando *p* está próximo a los bordes del intervalo [0,1].

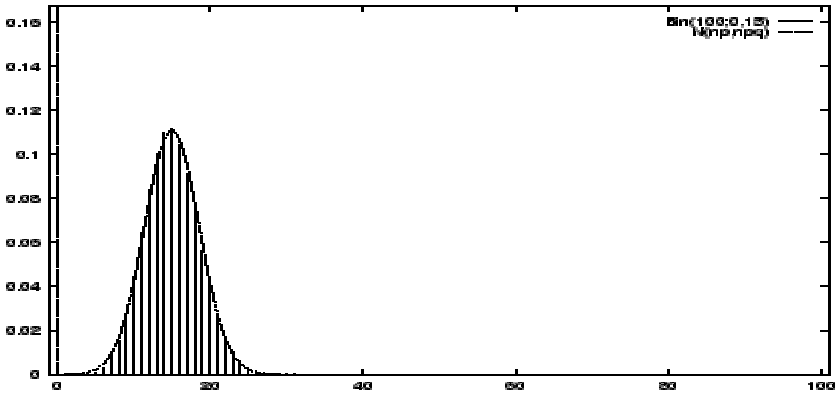
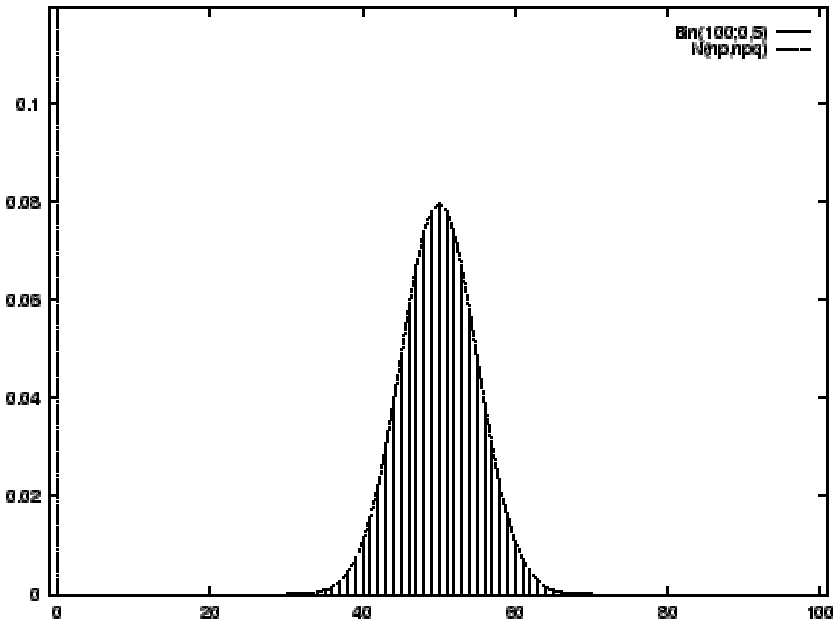


Figura 6: La misma comparación que en la figura anterior, pero realizada con parámetros con los que damos la aproximación normal de la binomial es mejor.



Ejemplo

Durante cierta epidemia de diarrea viral bovina, enferma el 30% de la una engorda. En un corral con 200 toretes, ¿cuál es la probabilidad de que al menos 40 padezcan la enfermedad? Calcular la probabilidad de que haya 60 toretes con diarrea viral bovina.

Solución: La v.a. que contabiliza el número de toretes que padece la enfermedad es:

$$X \rightsquigarrow \mathbf{B}(n = 200, p = 0,3)$$

cuya media es $\mu = n \cdot p = 60$ y su varianza es $\sigma^2 = npq = 42$. Realizar los cálculos con la ley binomial es muy engorroso, ya que intervienen números combinatorios de gran tamaño, y potencias muy elevadas. Por ello utilizamos la aproximación normal de *X*, teniendo en cuenta que se verifican las condiciones necesarias para que el error sea aceptable:

$$X \rightsquigarrow \mathbf{B}(n,p) \quad \text{donde} \quad \begin{cases} n = 200 > 30 \\ np = 60 > 4 \\ nq = 140 > 4 \end{cases} \quad \Longrightarrow \quad X \approx X_N \rightsquigarrow \mathbf{N}\left(\mu = 60, \sigma^2 = 42\right)$$

Así aproximando la v.a. discreta binomial *X*, mediante la v.a. continua normal X_N tenemos:

$$\begin{aligned}
\mathcal{P}[X \leq 40] &\approx \mathcal{P}[X_N \leq 40] \\
\textbf{tipificando la aproximación} &= \mathcal{P}[\underbrace{\frac{X-60}{\sqrt{42}}}_{Z;\ \mathrm{N}(0,1)} \leq \frac{40-60}{\sqrt{42}}] \approx \mathcal{P}[Z \leq -3'09] \\
\textbf{por simetría} &= \mathcal{P}[Z \geq 3'09] \\
\textbf{por el suceso contrario} &= 1 - \mathcal{P}[Z \leq 3'09] \\
\textbf{buscando en la tabla 3} &= 0,999
\end{aligned}$$

También es necesario calcular P[X=60]. Esta probabilidad se calcula exactamente como:

$$\mathcal{P}[X = 60] = \binom{200}{60} p^{60} q^{140}$$

Dada la dificultad numérica para calcular esa cantidad, y como la distribución binomial no está habitualmente tabulada hasta valores tan altos, vamos a utilizar su aproximación normal, X_N. Pero hay que prestar atención al hecho de que X_N es una v.a. continua, y por tanto la probabilidad de cualquier punto es cero. En particular,

$$\mathcal{P}[X_N = 60] = 0 \implies \mathcal{P}[X = 60] \approx 0$$

lo que ha de ser interpretado como un error de aproximación. Hay métodos más aproximados para calcular la probabilidad buscada. Por ejemplo, podemos aproximar P[X=60] por el valor de la función de densidad de X_N en ese punto (es en el único sentido en que se puede entender la función de densidad de la normal como una aproximación de una probabilidad). Así:

$$\mathcal{P}[X = 60] \approx f_{X_N}(60) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{60-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{0} = 0,063$$

Por último, otra posibilidad es considerar un intervalo de longitud 1 centrado en el valor 60 del que deseamos hallar su probabilidad y hacer:

$$\mathcal{P}[X=60] \approx \mathcal{P}[59'5 \leq X_N \leq 60,5] \approx \mathcal{P}[\underbrace{-0,08 \leq Z \leq 0,08}_{\text{simetría}}] = 0,0638$$

Ejemplo

Según un estudio, el peso al destete de becerros Hereford en cierto rancho es una v.a. *X*, que podemos considerar que se distribuye según una ley gaussiana de valor esperado μ = 175 kg y desviación típica σ = 10 kg. Dar un intervalo para el que tengamos asegurado que el 50% de los becerros del rancho comprendidos en él.

Solución: Tenemos que

$$X \rightsquigarrow \mathbf{N}\left(\mu = 175, \sigma^2 = 10^2\right)$$

.

Si buscamos un intervalo donde estar seguros de que el 50% de los becerros sus pesos comprendidos en él hay varias estrategias posibles:

1.- Podemos tomar el porciento 50, ya que este valor deja por debajo suya a la mitad, 0,5, de la masa de probabilidad. Este valor, $x_{0,5}$, se definiría como:

$$\int_{-\infty}^{x_{0,5}} f(t)\,dt = 0,5 \iff \mathcal{P}[X \leq x_{0,5}] = 0,5$$

$$\text{tipificando} \iff \mathcal{P}[Z \leq z_{0,5}] = 0,5$$

donde

$$Z = \frac{X-\mu}{\sigma} = \frac{X-175}{10} \rightsquigarrow \mathbf{N}(0,1)$$

$$z_{0,5} = \frac{x_{0,5}-\mu}{\sigma} = \frac{x_{0,5}-175}{10}$$

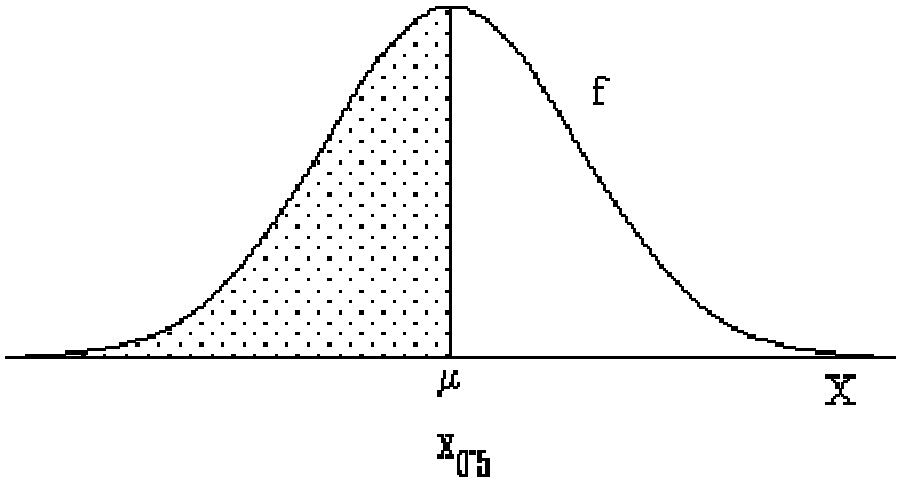
El valor $z_{0,5}$ lo podemos buscar en la tabla 3 (distribución N(0.1) y se obtiene

$$z_{0,5} = 0 \implies x_{0,5} = 175 + 10 \cdot z_{0,05} = 175$$

Por tanto podemos decir que la mitad de la población tiene un peso inferior a $X_{0.5}$ = 175 kg. Este resultado era de esperar, ya que en la distribución es simétrica y habrá una mitad de individuos con un peso inferior a la media y otro con un peso superior (figura 7). Esto puede escribirse como:

> El 50% de la población tiene un peso comprendido en el intervalo (-∞, 175].

Figura 7: Intervalo donde tenemos asegurado que el 50% de la población tiene un peso comprendido en él. Como se observa, no es un tamaño óptimo, en el sentido de que el intervalo es demasiado grande (longitud infinita a la izquierda).



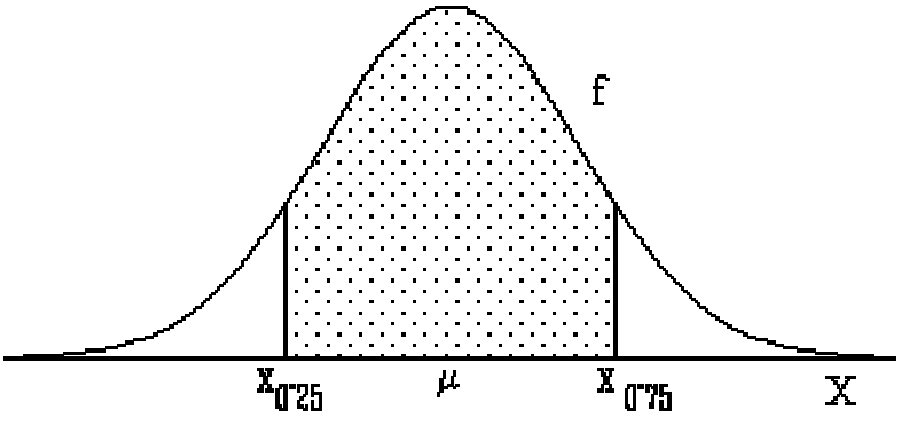
2.- Análogamente podemos considerar el por ciento 50, y tomar como intervalo aquellas alturas que lo superan. Por las mismas razones que en el problema anterior, podremos decir:

El 50% de la población tiene una altura comprendida en el intervalo [175, + ∞).

3.- Los anteriores intervalos, aún dando un resultado correcto, no son satisfactorios en el sentido de que son muy grandes, y no tienen en cuenta la simetría de la distribución normal para tomar un intervalo cuyo centro sea μ. Vamos a utilizar entonces otra técnica que nos permita calcular el intervalo centrado en la media, y que además será el más pequeño posible que contenga al 50% de la población.

Para ello observamos que la mayor parte de probabilidad está concentrada siempre alrededor de la media en las leyes gaussianas. Entonces podemos tomar un intervalo que contenga un 25% de probabilidad del lado izquierdo más próximo a la media, y un 25% del derecho (figura 8).

Figura 8: Intervalo donde tenemos asegurado que el 50% de la población tiene un peso comprendido en él. En este caso el intervalo es más pequeño que el anterior y está centrado en μ.



Esto se puede describir como el intervalo

$$[x_{0,25}\ ,\ x_{0,75}]$$

donde $x_{0,25}$ es el valor que deja por debajo de sí al 25% de la masa de probabilidad y $x_{0,75}$ el que lo deja por encima (o lo que es lo mismo, el que deja por debajo al 75% de las observaciones). Del mismo modo que antes estos valores pueden ser buscados en una tabla de la distribución normal, tipificando en primera instancia para destipificar después:

$$\int_{-\infty}^{x_{0,75}} f(t)\,dt = 0,75 \iff \mathcal{P}[X \leq x_{0,75}] = 0,75$$

$$\textbf{tipificando} \iff \mathcal{P}[Z \leq z_{0,75}] = 0,75$$

donde

$$Z = \frac{X-\mu}{\sigma} = \frac{X-175}{10} \rightsquigarrow \mathbf{N}(0,1)$$

$$z_{0,75} = \frac{x_{0,75}-\mu}{\sigma} = \frac{x_{0,75}-175}{10}$$

En una tabla encontramos el valor $z_{0,75}$, y se destipifica:

$$z_{0,75} = 0,675 \implies x_{0,75} = 175 + 10 \cdot z_{0,75} = 181'75$$

Análogamente se calcularía

$$\int_{-\infty}^{x_{0,25}} f(t)\,dt = 0,25 \iff \mathcal{P}[X \leq x_{0,25}] = 0,25$$

$$\textbf{tipificando} \iff \mathcal{P}[Z \leq z_{0,25}] = 0,25$$

donde

$$Z = \frac{X-\mu}{\sigma} = \frac{X-175}{10} \rightsquigarrow \mathbf{N}(0,1)$$

$$z_{0,25} = \frac{x_{0,25}-\mu}{\sigma} = \frac{x_{0,25}-175}{10}$$

Por la simetría de la distribución normal con respecto al origen, tenemos que $z_{0,25}$= - $z_{0,75}$.Luego

$$z_{0,25} = -0,675 \implies x_{0,75} = 175 + 10 \cdot z_{0,25} = 168,25$$

En conclusión:

El 50% de la población tiene un peso comprendido en el intervalo [168,25,181,75].

De entre los tres intervalos que se han calculado el que tiene más interés es el último, ya que es simétrico con respecto a la media, y es el más pequeño de todos los posibles (más preciso).

INTERVALOS DE CONFIANZA PARA LA DISTRIBUCIÓN NORMAL

Dada una variable aleatoria de distribución gaussiana $X \sim N(\mu,\sigma^2)$,nos intere--samos en primer lugar, en calcular intervalos de confianza para sus dos parámetros, μ y σ^2 . Enseguida se muestra un resumen de las situaciones consideradas.

Intervalo para la Media si se Conoce la Varianza

Este no es un caso práctico (no se puede conocer σ^2 sin conocer previamente μ), pero sirve para introducirnos en el problema de la estimación confidencial de la media.

Intervalos de Confianza para la Media (caso general)

Este se trata del caso con verdadero interés práctico. Por ejemplo sirve para estimar intervalos que contenga la media del colesterol en sangre en una población, la altura, el peso, etc, cuando disponemos de una muestra de la variable.

Intervalo de Confianza para la Varianza

Éste es otro caso de interés en las aplicaciones. El objetivo es calcular un intervalo de confianza para σ^2, cuando sólo se dispone de una muestra.

Estimación del tamaño muestral

La utilidad consiste en decidir cuál deberá ser el tamaño necesario de una muestra para obtener intervalos de confianza para una media, con precisión y significación dadas de antemano. Para que esto sea posible es necesario poseer cierta información previa, que se obtiene a partir de las denominadas **muestras piloto**.

Enseguida se considera caso en que se tiene dos poblaciones donde cada una sigue su propia ley de distribución $N(\mu_1,\sigma^2_1)$,y $N(\mu_2,\sigma^2_2)$. Los problemas asociados a este caso son:

Diferencia de medias homocedáticas

Se realiza el cálculo del intervalo de confianza suponiendo que ambas variables tienen la misma varianza, es decir **son homocedáticas.** En la práctica se usa este cálculo, cuando ambas variables tienen parecida dispersión.

Diferencia de medias (caso general)

Es el mismo caso que el anterior, pero se realiza cuando se observa que hay diferencia notable en la dispersión de ambas variables.

Intervalo para la Media si se Conoce la Varianza

Este caso que planteamos es más a nivel teórico que práctico: difícilmente vamos a poder conocer con exactitud σ^2 mientras que μ es desconocido. Sin embargo nos aproxima del modo más simple a la estimación confidencial de medias.

Para estimar μ el estadístico que mejor nos va a ayudar es $\overline{X}$, del que conocemos su ley de distribución:

$$\overline{X} \rightsquigarrow \underbrace{\mathbf{N}\left(\mu, \frac{\sigma^2}{n}\right)}_{\substack{\text{un parámetro} \\ \text{desconocido}}}$$

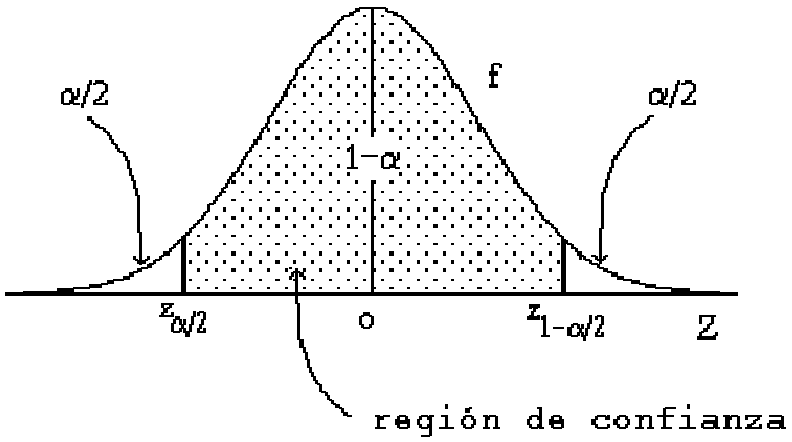
Esa ley de distribución depende de μ (desconocida). Lo más conveniente es hacer que la ley de distribución no dependa de ningún parámetro desconocido, para ello tipificamos:

$$Z = \underbrace{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}_{\substack{\text{par. desconocido} \\ + \\ \text{estimador} \\ + \\ \text{cosas conocidas}}} \rightsquigarrow \underbrace{\mathbf{N}(0,1)}_{\text{tabulada}}$$

Este es el modo en que haremos siempre la estimación puntual: buscaremos una relación en la que intervengan el parámetro desconocido junto con su estimador y de modo que estos se distribuyan según una ley de probabilidad que es bien conocida y a ser posible tabulada.

De este modo, fijado $\alpha \in (0,1)$ consideramos la v.a. $Z \sim N(0,1)$ y tomamos un intervalo que contenga una masa de probabilidad de 1-α. Este intervalo lo queremos tan pequeño como sea posible. Por ello lo mejor es tomarlo simétrico con respecto a la media (0), ya que allí es donde se acumula más masa (véase la figura 9). Así las dos colas de la distribución (zonas más alejadas de la media) se repartirán a partes iguales el resto de la masa de probabilidad, α.

Figura 9: La distribución N(0,1) y el intervalo más pequeño posible cuya probabilidad es 1- α. Por simetría, los cuantiles Z α/2 y $Z_{1-\alpha/2}$ sólo difieren en el signo.



Ahora se precisa cómo calcular el intervalo de confianza:

- Sea Zα/2 el porcentaje 100 α/2 de *Z*, es decir, aquel valor de *IR* que deja por debajo de si la cantidad Z α/2de la masa de probabilidad de *Z*, es decir:

$$\mathcal{P}[Z \leq z_{\alpha/2}] = \frac{\alpha}{2}$$

- Sea Z_{1-} α/2 el porciento 100-(1-α/2), es decir,

$$\mathcal{P}[Z \leq z_{1-\alpha/2}] = 1 - \frac{\alpha}{2}$$

Es útil considerar en este punto la simetría de la distribución normal, y observar que los porcentuales anteriores son los mismos aunque con el signo cambiado:

$$z_{\alpha/2} = -z_{1-\alpha/2}$$

- El intervalo alrededor del origen que contiene la mayor parte de la masa (1-α) es el intervalo siguiente (Figura 10):

$$\left[z_{\alpha/2}, z_{1-\alpha/2}\right] = \left[-z_{1-\alpha/2}, z_{1-\alpha/2}\right]$$

lo que habitualmente escribiremos como:

$$|Z| \leq z_{1-\alpha/2}$$

- De este modo podemos afirmar que existe una probabilidad de 1-α de que al extraer una muestra aleatoria de la variable en estudio, ocurra:

$$\begin{aligned} |Z| \leq z_{1-\alpha/2} \quad &\Rightarrow \\ &\Rightarrow \quad \frac{\left|\overline{X} - \mu\right|}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2} \\ &\Rightarrow \quad \left|\overline{X} - \mu\right| \leq z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \end{aligned}$$

De este modo un intervalo de confianza al nivel 1-α para la esperanza de una normal de varianza conocida es el comprendido entre los valores

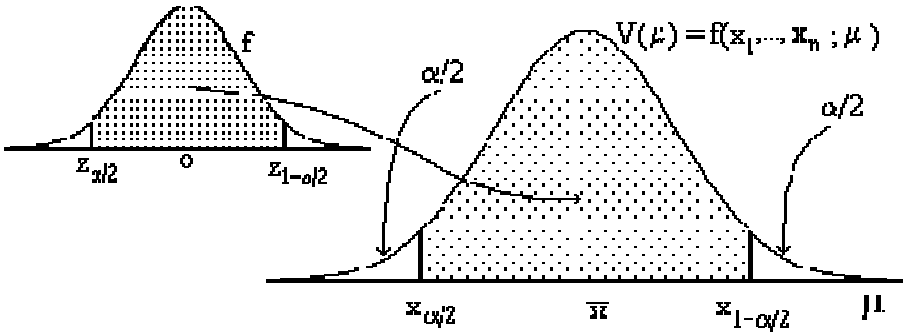
$$x_{\alpha/2} = \overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$x_{1-\alpha/2} = \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

La forma habitual de escribir este intervalo está inspirada en la 11 :

$$\mu = \overline{X} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Figura 10: Intervalo de confianza para la media.



Intervalo para la media (caso general)

Como hemos mencionado, los casos anteriores se presentarán poco en la práctica, ya que lo usual es que sobre una población quizás podamos conocer si se distribuye normalmente, pero el valor exacto de los parámetros μ y α^2 no son conocidos. De ahí el interés en buscar intervalos de confianza para ellos.

El problema que tenemos en este caso es más complicado que el anterior, pues no es tan sencillo eliminar los dos parámetros a la vez. Para ello nos vamos a ayudar de lo siguiente:

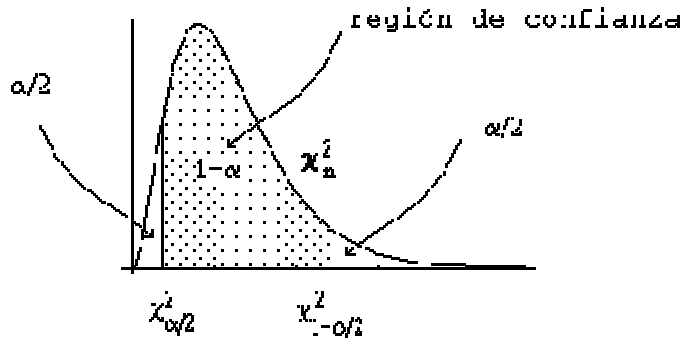
$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathbf{N}(0,1)$$

Por el teorema de Cochran se sabe por otro lado que:

$$\chi^2_{n-1} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

y que además estas dos últimas distribuciones son independientes. A partir de estas relaciones podemos construir una distribución t de Student con *n*-1 grados de libertad (figura11):

Figura 11: La distribución t_n es algo diferente a N(0,1) cuando *n* es pequeño, pero conforme éste aumenta, ambas distribuciones se aproximan.



$$T_{n-1} = \frac{Z}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}}$$

$$= \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}}}$$

Simplificando la expresión anterior tenemos:

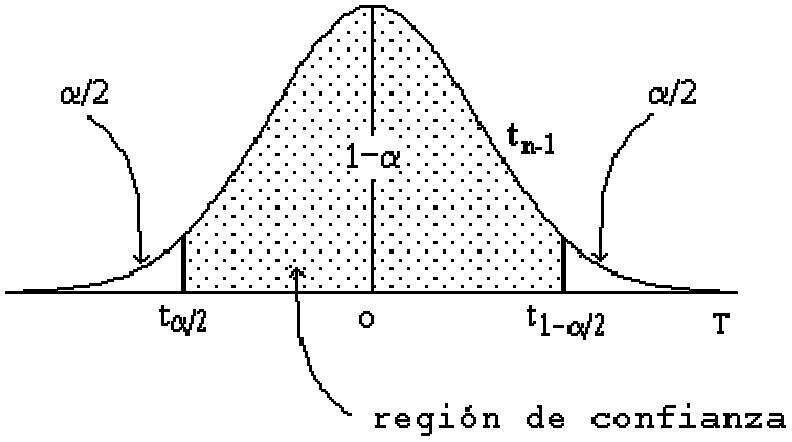
$$\boxed{T_{n-1} = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}} \rightsquigarrow t_{n-1}}$$

Dado el nivel de significación 1-α buscamos en una tabla de t_{n-1} el porcentual 100-(1-α/2), $t_{n-1,1-\alpha/2}$, el cual deja por encima de si la cantidad α/2 de la masa de probabilidad (figura 12). Por simetría de la distribución de Student se tiene que

$$t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2},$$

luego

Figura 12: La distribución de Student tiene las mismas propiedades de simetría que la normal tipificada.



$$\begin{cases} \mathcal{P}[T_{n-1} > t_{n-1,1-\alpha/2}] = \frac{\alpha}{2} \\ \mathcal{P}[T_{n-1} < -t_{n-1,1-\alpha/2}] = \frac{\alpha}{2} \end{cases} \iff \mathcal{P}[|T_{n-1}| \leq t_{n-1,1-\alpha/2}] = 1-\alpha$$

El intervalo de confianza se obtiene a partir del siguiente cálculo:

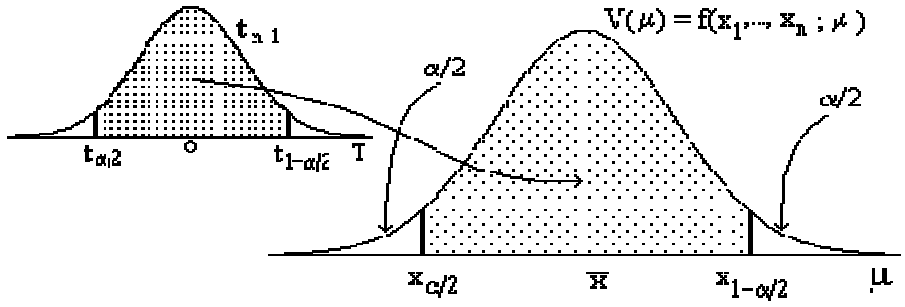
$$|T_{n-1}| \leq t_{n-1,1-\alpha/2} \Rightarrow \frac{|\overline{X}-\mu|}{\hat{S}/\sqrt{n}} \leq t_{n-1,1-\alpha/2}$$

$$\Rightarrow |Xmed-\mu| \leq t_{n-1,1-\alpha/2} \cdot \hat{S}/\sqrt{n}$$

Es decir, el intervalo de confianza al nivel 1-α para la esperanza de una distribución gaussiana cuando sus parámetros son desconocidos es:

$$\mu = \overline{X} \pm t_{n-1,1-\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Figura 13: Intervalo de confianza para μ cuando σ^2 es desconocido (caso general).



Al igual que en el caso del cálculo del intervalo de confianza para μ cuando σ^2 es conocido, podemos en el caso σ^2 desconocido, utilizar la función de verosimilitud (figura 15) para representarlo geométricamente. En este caso se usa la notación:

$$x_{\alpha/2} = \overline{x} - t_{n-1,1-\alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}$$

$$x_{1-\alpha/2} = \overline{x} + t_{n-1,1-\alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}$$

Ejemplo

Se quiere estimar un intervalo de confianza al nivel de significación α=0.05 para el peso medio al destete μ de becerros hereford. En principio sólo sabemos que la distribución de las alturas es una v.a. *X* de distribución normal. Para ello se toma una muestra de *n*=25 becerros y se obtiene

$\overline{X}$ = 170 kg

S = 10 kg

Solución:

$$S = 10 \Longrightarrow \hat{S} = S\sqrt{\frac{n}{n-1}} = 10\sqrt{\frac{25}{24}} = 10'206$$

Si queremos estimar un intervalo de confianza para μ, es conveniente utilizar el estadístico

$$T = \frac{\overline{x} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \rightsquigarrow \mathbf{t}_{n-1}$$

y tomar como intervalo de confianza aquella región en la que

$$|T| \leq t_{n-1;1-\alpha/2}$$

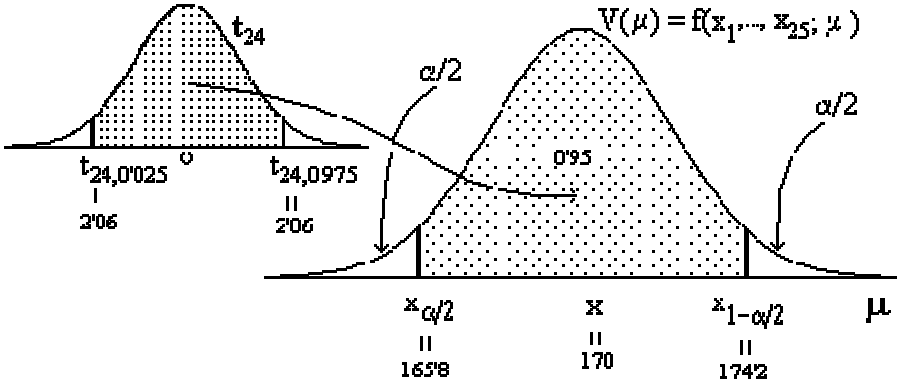
es decir,

$$\left| \frac{170 - \mu}{\frac{10,206}{\sqrt{25}}} \right| \leq t_{24;0,975} = 2,06 \Longrightarrow \mu = 170 \pm 2,06 \cdot \frac{10,206}{5} = 170 \pm 4,204$$

o dicho de forma más precisa: Con un nivel de confianza del 95% podemos decir que la media poblacional está en el intervalo siguiente (Figura 15)

$$\mu \in [165,796\,;\,174,204]$$

Figura 14: Cálculo del intervalo de confianza para la media usando para ello la distribución de Student y la función de verosimilitud asociada, la ..cual .ésta tiene su máximo en $\overline{X}$, ya que esta estimación puntual de μ es la *máximo verosímil*.



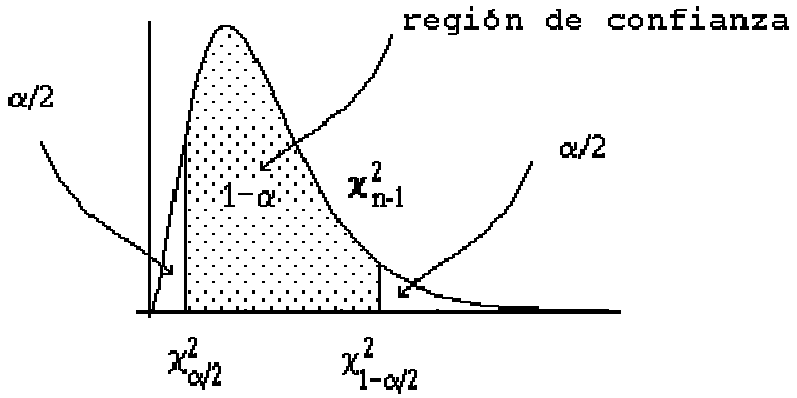
Intervalo de Confianza para la Varianza

Para estimar un intervalo de confianza para la varianza, nos ayudaremos de la siguiente propiedad de la distribución X^2

$$\chi^2_{n-1} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{(n-1)\,\hat{S}^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

Consideremos dos cuantiles de esta distribución que nos dejen una probabilidad 1-α en la ``zona central'' de la distribución (cf. Figura 16):

Figura 15: Cuantiles de la distribución X^2_{n-1}



$$\begin{cases} \mathcal{P}\left[\chi^2_{n-1} < \chi^2_{n-1,\alpha/2}\right] = \frac{\alpha}{2} \\ \mathcal{P}\left[\chi^2_{n-1} > \chi^2_{n-1,1-\alpha/2}\right] = \frac{\alpha}{2} \end{cases} \Longrightarrow \mathcal{P}\left[\chi^2_{n-1,\alpha/2} \leq \chi^2_{n-1} \leq \chi^2_{n-1,1-\alpha/2}\right] = 1-\alpha$$

Entonces un intervalo de confianza al nivel 1-α para la varianza de una distribución gaussiana (cuyos parámetros desconocemos) lo obtenemos teniendo en cuenta que existe una probabilidad 1-α de que:

$$\chi^2_{n-1,\alpha/2} \leq \chi^2_{n-1} \leq \chi^2_{n-1,1-\alpha/2} \Rightarrow$$

$$\Rightarrow \chi^2_{n-1,\alpha/2} \leq \frac{(n-1)\hat{S}^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}$$

$$\Rightarrow \frac{(n-1)\hat{S}^2}{\chi^2_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)\hat{S}^2}{\chi^2_{n-1,\alpha/2}}$$

Por tanto el intervalo que buscamos es

$$\sigma^2 \in \left[\frac{(n-1)\hat{S}^2}{\chi^2_{n-1,1-\alpha/2}} \ , \ \frac{(n-1)\hat{S}^2}{\chi^2_{n-1,\alpha/2}}\right]$$

(Universidad de Malaga, 1999)

Ejemplo

En el se estudió el peso al destete de becerros Hereoford obteniéndose en una muestra de tamaño 25 los siguientes valores:

$\overline{X}$ = 170 kg

S = 10 kg

Calcular un intervalo de confianza con α=0.05 para la varianza σ^2 de la altura de los individuos de la ciudad.

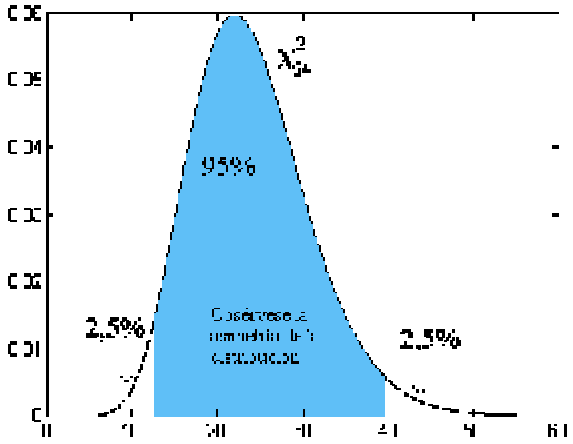
Solución:

Para estimar un intervalo de confianza para α^2 (varianza poblacional) el estadístico es:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

Entonces el intervalo de confianza que buscamos lo obtenemos mediante (cf. Figura 16)

Figura 16: Percentiles del 2,5% y del 97,5%
para la distribución X^2_{24}.



$$\chi^2_{n-1,\alpha/2} \leq \chi^2 \leq \chi_{n-1,1-\alpha^2} \iff \chi^2_{24;0,025} = 12,4 \leq \frac{24 \cdot 10,206^2}{\sigma^2} \leq \chi_{24;0,075} = 39'4$$

$$\iff \sigma^2 \in [63,45; 201,60]$$

Por tanto, para el valor poblacional de la desviación típica tenemos que

$$7,96 \leq \sigma \leq 14,199$$

con una confianza del 95%, que por supuesto contiene a las estimaciones puntuales *S*=10 y Ŝ=10.206 calculados sobre la muestra.

LITERATURA CITADA.

Bioestadística: Métodos y Aplicaciones. Universidad de Málaga. http://www.bioestadistica.uma.es/libro/

Castillo P.J, J.G.Arias. 1998. Estadística inferencial básica. Grupo editorial Ibero América. México.

Cochran. G. William. 1980. Diseños experimentales. Editorial Trillas. México

Infante G. S. 1997. Métodos estadísticos. Editorial Trillas. México.

Kreyszig Erwin. 1979. Introducción a la estadística matemática. Editorial Limusa. México

Montgomery D. C.1991.Diseño y análisis de experimentos. Grupo editorial Iberoamérica. México.

Ostle, B. 1965. Estadística aplicada. Primera edición. Editorial Limusa. México.

Rodríguez, D.A.J.M. 1991. Métodos de Investigación Pecuaria. 1ª Edición. Editorial Trillas. México.

Steel, R.G.D. 1986. Bioestadistica: Principios y Procedimientos. 1ª Edición. Editorial McGraw-Hill. México.

Universidad de Malaga. 1999. Bioestadística. Manual de la Universidad de Malaga. Bioestadística, Métodos y aplicaciones. ISBN: 847496-653-1

LITERATURA REVISADA

Bioestadística: Métodos y Aplicaciones. Universidad de Málaga. http://www.bioestadistica.uma.es/libro/

Castillo P.J, J.G.Arias. 1998. Estadística inferencial básica. Grupo editorial Ibero América. México.

Cochran. G. William. 1980. Diseños experimentales. Editorial Trillas. México

Infante G. S. 1997. Métodos estadísticos. Editorial Trillas. México.

Kreyszig Erwin. 1979. Introducción a la estadística matemática. Editorial Limusa. México

Montgomery D. C.1991.Diseño y análisis de experimentos. Grupo editorial Iberoamérica. México.

Ostle, B. 1965. Estadística aplicada. Primera edición. Editorial Limusa. México.

Rodríguez del A. J. 1991. Métodos de investigación pecuaria. Editorial Trillas. México.

Snedecor W. George, W. G. Cochran. 1979 métodos estadísticos. Editorial Continental. México.

Steel G.D Robert, J. H. Torrie.1981. 2ª. Principles and procedures of statistics a biometrical approach. 2ª. Ed. Editorial Mc Graw-Hill. USA.

Walpole, R. E.1992. Probabilidad y estadística. Cuarta Edición. Editorial Mc GrawHill. México.

. MEDIDAS DE CENTRALIDAD Y DISPERSIÓN

INTRODUCCIÓN

En el trabajo estadístico es importante saber cuándo estamos tratando con una población completa de observaciones, o con una muestra de observaciones seleccionadas de una población especificada.Una población se puede definir como la totalidad de valores posibles (mediciones o conteos) de una característica particular de un grupo especificado de objetos. Tal grupo especificado de objetos se le conoce como un universo. Obviamente, un universo puede tener varias poblaciones asociadas con él.

Estos ejemplos son suficientes para indicar al lector de la importancia de definir claramente la población en investigación. El concepto de muestra, corresponde a una población, seleccionada de acuerdo con una regla o una estrategia. Las cosas importantes que se deben saber son:

a) Que estamos tratando con una muestra.

b) Qué población ha sido muestreada.

Si tratamos con toda la población, nuestro trabajo estadístico será parcialmente descriptivo. Por el contrario, si tratamos con una muestra el trabajo estadístico no únicamente describe a la muestra sino que también proporciona información respecto a la población muestreada.

En la práctica se han encontrado varias clases o tipos de muestras. Las características que distinguen a un tipo de otro son:

- La manera de obtención de la muestra.
- El número de variables considerables.

- El fin para que fue extraída la muestra.

Las dos últimas características se obtienen fácilmente en cualquier situación practica, aunque la última no es enunciada claramente y tal vez, olvida. La manera de obtener la muestra es muy importante y será discutida posteriormente.

Las muestras pueden agruparse en dos grandes clases, cuando se considera su método de selección, a saber, las que seleccionan por criterios y las que se seleccionan por medio de un mecanismo casual. Las muestras elegidas de acuerdo con el mecanismo casual, son llamadas muestras de probabilidad, si cada elemento de población tiene una probabilidad conocida de pertenecer a la muestra. En particular, si cada elemento tiene la misma probabilidad de pertenecer a la muestra, entonces ésta es conocida como una muestra al azar.
Las muestras al azar se prefieren a las muestras elegidas. Una buena muestra es aquella que a partir de la cual puede hacerse generalizaciones.

Para generalizar, de la muestra de población, se necesita estar capacitados para deducir, a partir de cualquiera de las suposiciones respecto a la población, cuándo la muestra observa está dentro del rango de variación del muestreo que puede ocurrir para dicha población, bajo el método dado de muestreo. Tales deducciones pueden hacerse, si y sólo si, se aplican las leyes de la probabilidad matemática. El objetivo de la naturaleza al azar de este tipo de muestreo. Tales deducciones pueden hacerse, si y sólo si, se aplican las leyes de la probabilidad matemática. El objetivo de la naturaleza al azar de este tipo de muestreo, es asegurar que esas leyes son aplicables.

El muestreo de diferentes poblaciones puede hacer de diferentes maneras:

- Una muestra al azar puede extraerse de una población especificada mediante una función continua de densidad de probabilidad. En este caso no puede presentarse el problema de muestreo con o sin reemplazamiento.

- Una muestra al azar puede extraerse de una población infinita especificada, mediante una función discreta de densidad de probabilidad donde el muestreo se efectúa con reemplazamiento. El muestreo con reemplazamiento tiene por objeto hacer infinita a la población.

El conocer las características de una muestra es evidentemente necesario pero también es necesario analizar a través de medidas de tendencia central y dispersión.

Medidas de tendencia central.

Tendencia Central

- Las medidas de tendencia central son puntos en una distribución, los valores medios o centrales de ésta y nos ayudan a ubicarla dentro de la escala de medición.

Las principales medidas de tendencia central son tres: moda, mediana y media.

La media

La **media aritmética** de una variable estadística es la suma de todos sus posibles valores, ponderada por las frecuencias de los mismos. Es decir, si la tabla de valores de una variable *X* es

X	n_i	f_i
x_1	n_1	f_1
...	...	...
x_k	n_k	f_k

la media es el valor que podemos escribir de las siguientes formas equivalentes:

$$\begin{aligned} \overline{x} &= x_1 f_1 + \ldots + x_k f_k \\ &= \frac{1}{n}\left(x_1 n_1 + \ldots x_k n_k\right) \\ &= \frac{1}{n}\sum_{i=1}^{k} x_i n_i \end{aligned}$$

Si los datos no están ordenados en una tabla, entonces

$$\overline{x} = \frac{x_1 + \ldots + x_n}{n}$$

Observación

Hemos supuesto implícitamente en la definición de media que tratábamos con una variable *X* discreta. Si la variable es continua tendremos que cambiar los valores de x_i por las marcas de clase correspondientes. En general, la media aritmética obtenida a partir de las marcas de clase c_i, diferirá de la media obtenida con los valores reales, x_i. Es decir, habrá una perdida de precisión que será tanto mayor cuanto mayor sea la diferencia entre los valores reales y las marcas de clase, o sea, cuanto mayores sean las longitudes a_i, de los intervalos.

Proposición

La suma de las *diferencias de la variable con respecto a la media* es nula, es decir,

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

Demostración.

Basta desarrollar la sumatoria para obtener

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + \ldots + (x_n - \overline{x}) = (x_1 + \ldots + x_n) - n\,\overline{x} = n\,\overline{x} - n\,\overline{x} = 0$$

Este resultado nos indica que el error cometido al aproximar un valor cualquiera de la variable, por ejemplo x_1, mediante el valor central $\overline{x}$, es compensado por los demás errores:

$$\textbf{Error aprox. de } x_1 \qquad \equiv \qquad x_1 - \overline{x} = \sum_{i=2}^{n}(x_i - \overline{x})$$

Si los errores se consideran con signo positivo, en este caso no pueden compensarse. Esto ocurre si tomamos como medida de error alguna de las siguientes:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \geq 0 \qquad \textbf{Error cuadrático}$$

$$\sum_{i=1}^{n} \mid x_i - \overline{x} \mid \geq 0 \qquad \textbf{Error absoluto}$$

$$\max_{i=1,\ldots,n} \mid x_i - \overline{x} \mid \geq 0 \qquad \textbf{Error máximo}$$

que son cantidades *estrictamente* positivas si algún $x_i \neq \overline{x}$.

Ejemplo

Obtener las desviaciones con respecto a la media en la siguiente distribución y comprobar que su suma es cero. Los datos corresponden a la calificación de condición corporal de 20 vacas lecheras con mas de 100 días en leche.

l_{i-1} - l_i	n_i
2.0 – 2.5	2
2.5 – 3.0	4
3.0 – 3.5	8
3.5 – 4.0	6

Solución:

l_{i-1} - l_i	n_i	x_i	$x_i\, n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})n_i$

2.0 – 2.5	2	2.25	4.5	- .95	- 1.9
2.5 – 3.0	4	2.75	11	- .45	- 1.8
3.0 – 3.5	8	3.25	26	+.05	+.4
3.5 – 4.0	6	3.75	22.5	+.55	+3.3
	n=20		$\sum x_i n_i = 3.2$		$\sum = 0$

La media aritmética es:

$$\overline{X} = \frac{1}{n}\sum x_i n_i = \frac{64}{20} = 3.2$$

Como se puede comprobar sumando los elementos de la última columna,

$$\sum (x_i - \overline{x}) \cdot n_i = 0$$

Proposición (Linealidad de la media)

$$Y = a + b\,X \Longrightarrow \overline{y} = a + b\,\overline{x}$$

Proposición

Dados *r* grupos con n_1, n_2, ..., n_r observaciones y siendo $\overline{x_1}\overline{x}_2,...,\overline{x}_r$ las respectivas medias de cada uno de ellos. Entonces la media de las $n = n_1 + ... + n_r$ observaciones es

$$\overline{x} = \frac{n_1\,\overline{x}_1 + \ldots + n_r\,\overline{x}_r}{n_1 + \ldots + n_r}$$

Demostración

Vamos a llamar x_{ij} a la *j*-ésima observación del grupo *i*; Entonces tenemos

$$\left.\begin{array}{lcccc} 1^{er}\ \text{grupo} & \longrightarrow & x_{11} & \ldots & x_{1n_1} \\ 2^{o}\ \text{grupo} & \longrightarrow & x_{21} & \ldots & x_{2n_2} \\ & \ldots & & & \\ r^{esimo}\ \text{grupo} & \longrightarrow & x_{r1} & \ldots & x_{rn_r} \end{array}\right\} \Longrightarrow \begin{array}{l} \overline{x}_1 = \left(\sum_{j=1}^{n_1} x_{ij}\right)/n_1 \\ \overline{x}_2 = \left(\sum_{j=1}^{n_2} x_{ij}\right)/n_2 \\ \ldots \\ \overline{x}_r = \left(\sum_{j=1}^{n_r} x_{ij}\right)/n_r \end{array}$$

Así, agrupando convenientemente las observaciones se llega a que

$$\begin{aligned}\overline{x} &= \frac{(x_{11}+\ldots+x_{1n_1})+(x_{22}+\ldots+x_{2n_2})+\ldots+(x_{r1}+\ldots+x_{rn_r})}{n_1+n_2+\ldots+n_r} \\ &= \frac{n_1\,\overline{x}_1+\ldots+n_r\,\overline{x}_r}{n}\end{aligned}$$

Observación

A pesar de las buenas propiedades que ofrece la media, ésta posee algunos inconvenientes:

- Uno de ellos es que es muy sensible a los valores extremos de la variable: ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección. En consecuencia,
- no es recomendable usar la media como medida central en las distribuciones muy asimétricas;
- Depende de la división en intervalos en el caso de variables continuas.

Cálculo abreviado

Se puede utilizar la linealidad de la media para simplificar las operaciones necesarias para su cálculo mediante un *cambio de origen* y de *unidad de medida*. El método consiste en lo siguiente:

1. Tomamos *a* un número que exprese aproximadamente el tipo de unidad con la que se trabaja. Por ejemplo, si las unidades que usamos son millones, tomamos *a* = 1.000.000.

2. Seleccionamos un punto cualquiera de la zona central de la tabla, x_0. Este punto jugará el papel de origen de referencia.

3. Cambiamos a la variable

$$Z = \frac{X - x_0}{a} \qquad \Longrightarrow \quad \overline{z} = \frac{\overline{x} - x_0}{a}$$
$$\Longrightarrow \quad \overline{x} = a\,\overline{z} + x_0$$

4. Construimos de este modo la tabla de la variable *Z*, para la que es más fácil calcular $\overline{X}$ directamente, y después se calcula $\overline{X}$ mediante la relación (2.2).

Medias generalizadas

En función del tipo de problema varias generalizaciones de la media pueden ser consideradas. He aquí algunas de ellas aplicadas a unas observaciones x_1, ..., x_n:

La media geométrica

$\overline{X}_y$, es la media de los logaritmos de los valores de la variable:

$$\log \overline{x}_g = \frac{\log x_1 + \ldots + \log x_n}{n}$$

Luego

$$\overline{x}_g = \sqrt[n]{x_1\, x_2 \ldots x_n}$$

Si los datos están agrupados en una tabla, entonces se tiene:

$$\overline{x}_g = \sqrt[n]{x_1^{n_1}\, x_2^{n_2} \ldots x_k^{n_k}}$$

La media armónica

$\overline{X}_a$, se define como el recíproco de la media aritmética de los recíprocos, es decir,

$$\frac{1}{\overline{x}_a} = \frac{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}{n}$$

Por tanto,

$$\overline{x}_a = \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}$$

La media cuadrática

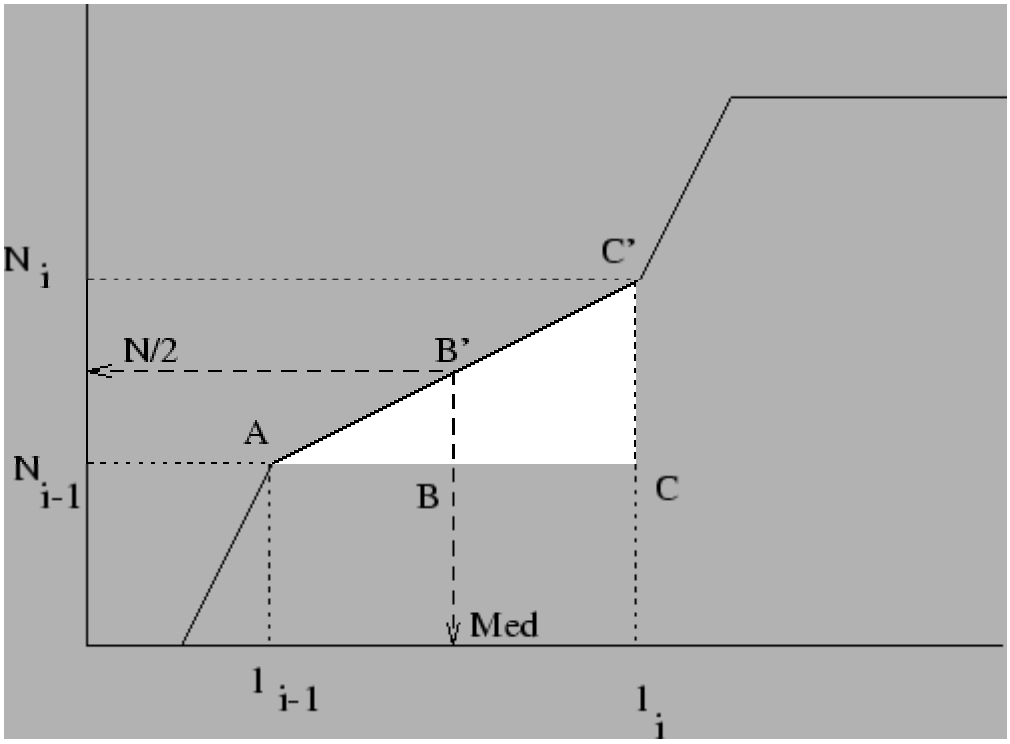
$\overline{X}_c$, es la raíz cuadrada de la media aritmética de los cuadrados:

$$\overline{x}_c = \sqrt{\frac{x_1^2 + \ldots + x_n^2}{n}}$$

La mediana

Consideramos una variable discreta *X* cuyas observaciones en una tabla estadística han sido ordenadas de menor a mayor. Llamaremos **mediana**, M_{ed} al primer valor de la variable que deja por debajo de sí al 50% de las observaciones. Por tanto, si *n* es el número de observaciones, la mediana corresponderá a la observación [*n*/2]+1, donde representamos por {.} parte entera de un número.

Figura 1: Cálculo geométrico de la mediana



En el caso de variables continuas, las clases vienen dadas por intervalos, y aquí la fórmula de la mediana se complica un poco más (pero no demasiado): Sea $(l_{i-1},l_i]$ el intervalo donde hemos encontrado que por debajo están el 50% de las observaciones. Entonces se obtiene la mediana a partir de las frecuencias absolutas acumuladas, mediante interpolación lineal (teorema de Thales) como sigue (figura 1):

$$\frac{CC'}{AC} = \frac{BB'}{AB} \implies \frac{n_i}{a_i} = \frac{\frac{n}{2} - N_{i-1}}{M_{ed} - l_{i-1}}$$

$$\implies \boxed{M_{ed} = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i}$$

Observación

La relación (1) corresponde a definir para cada posible observación, $x\varepsilon\left(l_{j-1}, l_j\right]$, su frecuencia relativa acumulada, $F(x)$, por interpolación lineal entre los valores $F(l_{j-1}) = F_{j-1}$ y $F(l_j) = F_j$ de forma que

$$F(x) = F(l_{j-1}) + \frac{F(l_j) - F(l_{j-1})}{a_j}(x - l_{j-1})$$

De este modo, M_{ed} es el punto donde $F\left(M_{ed}\right) = \frac{1}{2}$. Esto equivale a decir que *la mediana divide al histograma en dos partes de áreas iguales a*1/2.

Observación

Entre las propiedades de la mediana, vamos a destacar las siguientes:

- Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas.
- Es de cálculo rápido y de interpretación sencilla.
- A diferencia de la media, la mediana de una variable discreta es siempre un valor de la variable que estudiamos (ej. La mediana de una variable *número de hijos* toma siempre valores enteros).
- Si una población está formada por 2 subpoblaciones de medianas M_{ed1} y M_{ed2}, sólo se puede afirmar que la mediana, M_{ed}, de la población está comprendida entre M_{ed1} y M_{ed2}

$$M_{ed1} \leq M_{ed} \leq M_{ed2}$$

- El mayor defecto de la mediana es que tiene unas propiedades matemáticas complicadas, lo que hace que sea muy difícil de utilizar en *inferencia estadística*.
- Es función de los intervalos escogidos.
- Puede ser calculada aunque el intervalo inferior o el superior no tenga límites.
- La suma de las diferencias de los valores absolutos de *n* puntuaciones respecto a su mediana es menor o igual que cualquier otro valor. Este es el equivalente al teorema de König (proposición 2.1) con respecto a la media, pero donde se considera como medida de dispersión a:

$$\sum_{i=1}^{n} | x_i - M_{ed} |$$

Ejemplo

Obtener la media aritmética y la mediana en la distribución adjunta. Determinar gráficamente cuál de los dos promedios es más significativo. Los datos corresponden a la edad en días del nacimiento al primer servicio de la crianza de un establo lechero.

l_{i-1} - l_i	n_i
0 - 10	60
10 - 20	80
20 - 30	30
30 - 100	20
100 - 500	10

Solución:

l_{i-1} - l_i	n_i	a_i	x_i	$x_i\, n_i$	N_i	n_i'
0 - 10	60	10	5	300	60	60
10 - 20	80	10	15	1.200	140	80
20 - 30	30	10	25	750	170	30
30 - 100	20	70	65	1.300	190	2,9
100 - 500	10	400	300	3.000	200	0,25
	n=200			$\sum x_i n_i = 6.550$		

La media aritmética es:

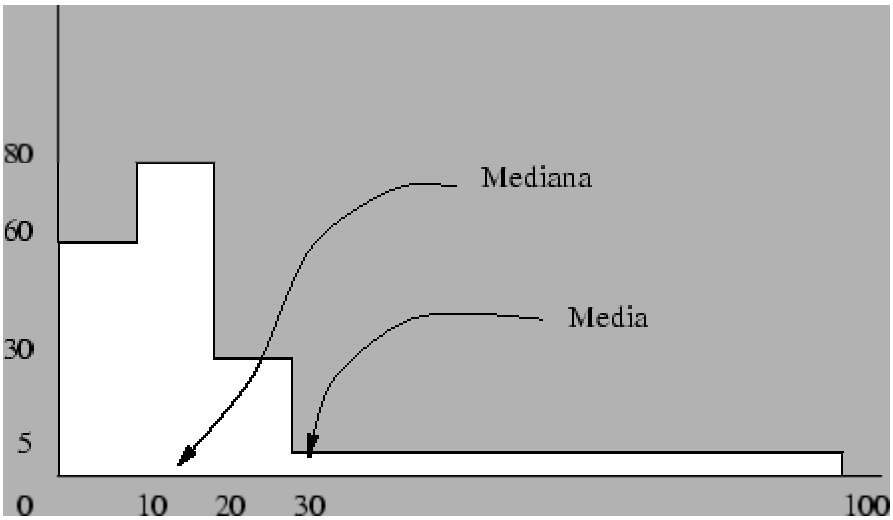
$$\overline{x} = \frac{1}{n}\sum x_i = \frac{6.550}{200} = 32,75 \ (días)$$

La primera frecuencia absoluta acumulada que supera el valor $n/2$=100 es N_i=140. Por ello el intervalo mediano es [10;20). Así:

$$M_{ed} = l_{i-1} + \frac{n/2 - N_{i-1}}{n_i} \cdot a_i = 10 + \frac{100-60}{80} \times 10 = 15 \ (días)$$

Para ver la representatividad de ambos promedios, realizamos el histograma de la figura 2, y observamos que dada la forma de la distribución, la mediana es más representativa que la media.

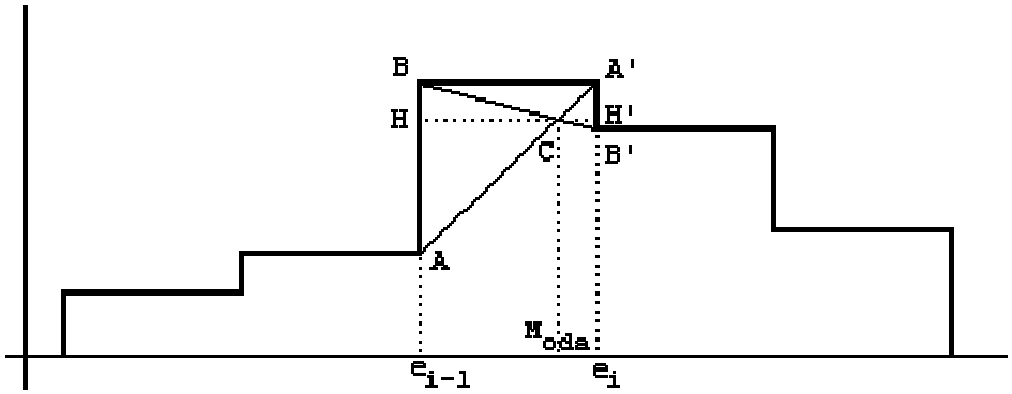
Figura 2: Para esta distribución de frecuencias es más representativo usar como estadístico de tendencia central la mediana que la media.



La moda

Llamaremos **moda** a cualquier máximo relativo de la distribución de frecuencias, es decir, cualquier valor de la variable que posea una frecuencia mayor que su anterior y su posterior.

Figura: Cálculo geométrico de la moda



Ejemplo

Obtener la Moda de los pesos al nacimiento de becerras Holstein:
37, 38, 41, 37, 40, 45, 44, 40, 39, 36, 40, 43, 40, 42,kg

Moda = 37, 40 kg

En el caso de variables continuas es más correcto hablar de *intervalos modales*. Una vez que este intervalo, (l_{i-1}, l_i], se ha obtenido, se utiliza la siguiente fórmula para calcular la moda, que está motivada en la figura 2.4:

En el caso de datos agrupados donde se ha construido una curva de frecuencias para ajustar los datos, la moda será el valor (o valores) de x correspondientes al máximo (o máximos) de la curva.

De una distribución de frecuencias o un histograma la moda puede obtenerse por:

$$\text{Moda} = L_1 + \frac{D_1}{D_1 + D_2} C$$

Donde:

L_1 = Límite real inferior de la clase modal (es decir, la clase que contiene la moda).

D_1 = Exceso de la frecuencia modal sobre la frecuencia de la clase continua inferior.

D_2 = Exceso de la frecuencia modal sobre la frecuencia de la clase continua superior.

C = Tamaño del intervalo de la clase modal.

Ejemplo.

En base a la siguiente distribución de frecuencias, encontrar la moda, de los siguientes limites del pesos de un destete (60 días) en kg de becerras Holstein.

Limite de las clases (kg)	Frecuencias
$52.5 < X \leq 57.5$	8
$57.5 < X \leq 62.5$	9
$62.5 < X \leq 67.5$	6
$67.5 < X \leq 72.5$	4
$72.5 < X \leq 77.5$	2
$77.5 < X \leq 82.5$	1
	30

$L_1 = 57.5$ $D_1 = 9 - 8 = 1$ $D_2 = 9 - 6 = 3$ $C = 5$

$$\text{Moda} = L_1 + \frac{D_1}{D_1 + D_2} C = 57.5 + \frac{1}{1+3}(5) = 57.5 + \frac{5}{4} = 57.5 + 1.25 = 58.75$$

Observación

De la moda destacamos las siguientes propiedades:

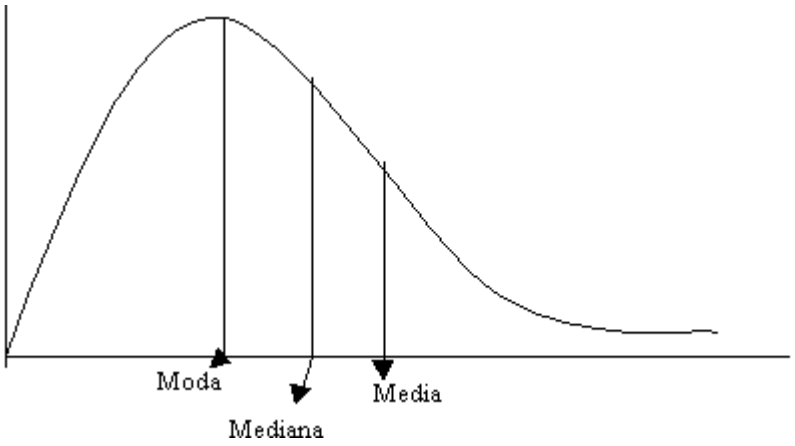
- Es muy fácil de calcular.
- Puede no ser única.
- Es función de los intervalos elegidos a través de su amplitud, número y límites de los mismos.
- Aunque el primero o el último de los intervalos no posean extremos inferior o superior respectivamente, la moda puede ser calculada.

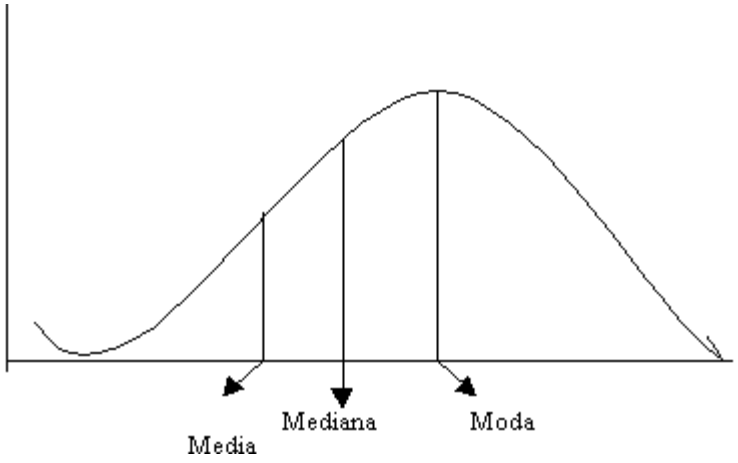
Relación entre media, mediana y moda

En el caso de distribuciones unimodales, la mediana está con frecuencia comprendida entre la media y la moda (incluso más cerca de la media).

En distribuciones que presentan cierta inclinación, es más aconsejable el uso de la mediana. Sin embargo en estudios relacionados con propósitos estadísticos y de inferencia suele ser más apta la media.

Veamos un ejemplo de cálculo de estas tres magnitudes.





Ejemplo

Consideramos una tabla estadística relativa a una variable continua, de la que nos dan los intervalos, las marcas de clase c_i, y las frecuencias absolutas, n_i, de las veces en que las cabras presentaron mastitis a largo de un periodo de lactancia.

Intervalos	c_i	n_i

0 -- 2	1	2
2 -- 4	3	1
4 -- 6	5	4
6 -- 8	7	3
8 - 10	9	2

Para calcular la media podemos añadir una columna con las cantidades $n_1 c_1$. La suma de los términos de esa columna dividida por n =12 es la media:

Intervalos	c_i	n_i	N_i	n_1c_1
0 -- 2	1	2	2	2
2 -- 4	3	1	3	3
4 -- 6	5	4	**7**	20
6 -- 8	7	3	10	21
8 - 10	9	2	12	18
	12		64	

$$\overline{X} = \frac{64}{12} = 5.\hat{3}$$

La **mediana** es el valor de la variable que deja por debajo de sí a la mitad de las n observaciones, es decir 6. Construimos la tabla de las frecuencias absolutas acumuladas, N_i, y vemos que eso ocurre en la modalidad tercera, es decir,

$$\begin{aligned} i &= 3 \qquad \textbf{Observación} \\ (l_{i-1}, l_i] &= (4;6] \qquad \textbf{Intervalo donde se encuentra la mediana} \\ M_{ed} &= l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i = 4 + \frac{\frac{12}{2} - 3}{4} \cdot 2 = 5,5 \in (l_{i-1}, l_i] \end{aligned}$$

Para el cálculo de la **moda**, lo primero es encontrar los intervalos modales, buscando los máximos relativos en la columna de las frecuencias absolutas, n_i.

Vemos que hay dos modas, correspondientes a las modalidades *i*=1, *i*=3. En el primer intervalo modal, ($l_{0,1}$]=(0,2], la moda se calcula como

$$M_{oda} = l_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} a_i = 0 + \frac{2-0}{(2-0)+(2-1)} 2 = 1,\widehat{3}$$

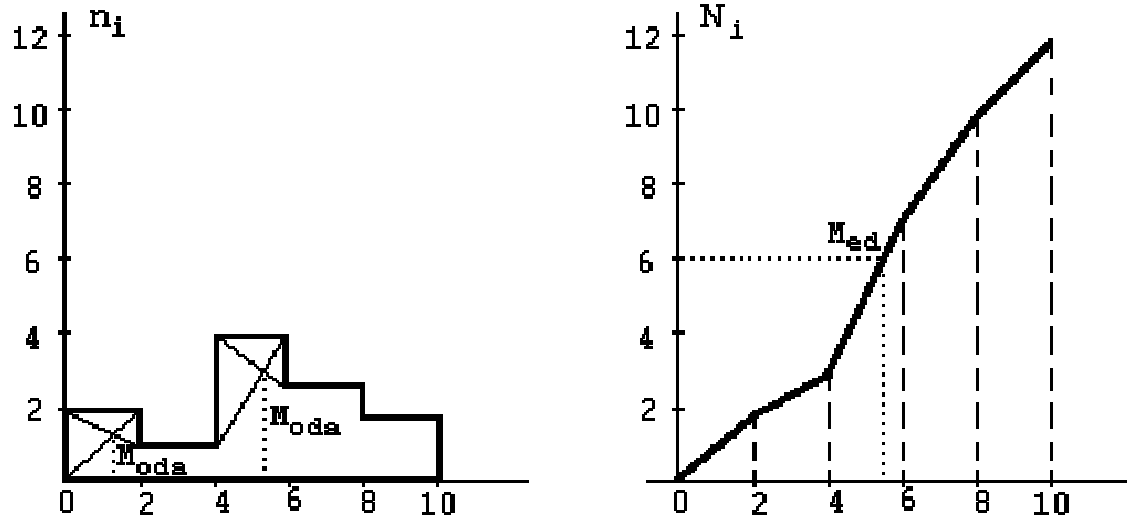
El segundo intervalo modal es (l_2,l_3]=(4;6], siendo la moda el punto perteneciente al mismo que se obtiene como:

$$M_{oda} = l_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} a_i = 4 + \frac{4-1}{(4-1)+(4-3)} 2 = 5,5$$

En este caso, como se ve en la figura 3, la moda no toma un valor único, sino el conjunto

$$M_{oda} = \{1,\widehat{3};\ 5,5\}$$

Figura 3: Diagramas diferencial e integral con cálculo geométrico de la moda y de la mediana de la variable.



Medidas de variabilidad o dispersión

Los estadísticos de tendencia central o posición nos indican donde se sitúa un grupo de puntuaciones. Los de variabilidad *o* dispersión nos indican si esas

puntuaciones o valores están próximas entre sí o si por el contrario están o muy dispersas.

Una medida razonable de la variabilidad podría ser la **amplitud** o **rango**, que se obtiene restando el valor más bajo de un conjunto de observaciones del valor más alto. Es fácil de calcular y sus unidades son las mismas que las de la variable, aunque posee varios inconvenientes:

- No utiliza todas las observaciones (sólo dos de ellas);
- Se puede ver muy afectada por alguna observación extrema;
- El rango aumenta con el número de observaciones, o bien se queda igual. En cualquier caso nunca disminuye.

En el transcurso de esta sección, veremos medidas de dispersión mejores que la anterior. Estas se determinan en función de la distancia entre las observaciones y algún estadístico de tendencia central.

Desviación media, D_m

Se define la **desviación media** como la media de las diferencias en valor absoluto de los valores de la variable a la media, es decir, si tenemos un conjunto de n observaciones, x_1, ..., x_n, entonces

$$D_m = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$$

Si los datos están agrupados en una tabla estadística es más sencillo usar la relación

$$D_m = \frac{1}{n}\sum_{i=1}^{k}|x_i - \overline{x}|\, n_i$$

Como se observa, la desviación media guarda las mismas dimensiones que las

observaciones. La suma de valores absolutos es relativamente sencilla de calcular, pero esta simplicidad tiene un inconveniente: Desde el punto de vista geométrico, la distancia que induce la desviación media en el espacio de observaciones no es la natural (no permite definir ángulos entre dos conjuntos de observaciones). Esto hace que sea muy engorroso trabajar con ella a la hora de hacer inferencia a la población.

Varianza y desviación estándar

Como forma de medir la dispersión de los datos hemos descartado:

- $\sum_{i=1}^{n}\left(x_1 - \bar{x}\right)$, pues sabemos que esa suma vale 0, ya que las desviaciones con respecto a la media se compensan al haber términos en esa suma que son de signos distintos.
- Para tener el mismo signo al sumar las desviaciones con respecto a la media podemos realizar la suma con valores absolutos. Esto nos lleva a la D_m, pero como hemos mencionado, tiene poco interés por las dificultades que presenta.

Si las desviaciones con respecto a la media las consideramos al cuadrado, $\left(x_i - \bar{x}\right)^2$, de nuevo obtenemos que todos los sumandos tienen el mismo signo (positivo). Esta es además la forma de medir la dispersión de los datos de forma que sus propiedades matemáticas son más fáciles de utilizar. Vamos a definir entonces dos estadísticos que serán *fundamentales* en el resto del curso: La *varianza* y la *desviación típica*.

La **varianza**, δ^2, se define como la media de las diferencias cuadráticas de *n* puntuaciones con respecto a su media aritmética, es decir $\delta^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

Para datos agrupados en tablas, usando las notaciones establecidas en los

capítulos anteriores, la varianza se puede escribir como $\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i$

Una fórmula equivalente para el cálculo de la varianza está basada en lo siguiente:

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\,x_i\,\overline{x} + \overline{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\,\overline{x}\,\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i)}_{=\overline{x}} + \frac{1}{n}n\,\overline{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\,\overline{x}^2 + \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 \end{aligned}$$

Con lo cual se tiene

$$\boxed{S^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}$$

Si los datos están agrupados en tablas, es evidente que

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}x_i^2\,n_i - \overline{x}^2$$

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en m^2). Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada. Por ello se define la **desviación típica**, δ, como

$$S = \sqrt{S^2}$$

Ejemplo

Calcular la varianza y desviación estándar de las siguientes días diarrea por animal que se presentaron en animales que están en el perdió critico (primeras 3 semanas de vida):

3,3,4,4,5 días

Solución: Para calcular dichas medidas de dispersión es necesario calcular previamente el valor con respecto al cual vamos a calcular las diferencias. Ésta es la media:

$$\overline{X} = \frac{(3+3+4+4+5)}{5} = 3.8 días$$

La varianza es:

$$\delta^2 = \frac{1}{n}\sum_{i=1}^{n} x^2{}_i - \bar{x}^2 = \frac{1}{5}\left(3^2+3^2+4^2+4^2+5^2\right) - 3.8 = 0.56 días$$

siendo la desviación típica su raíz cuadrada:

$$\delta = \sqrt{\delta^2} = \sqrt{0.6} = .748 días$$

Las siguientes propiedades de la varianza (respectivamente, desviación típica) son importantes a la hora de hacer un cambio de origen y escala a una variable. En primer lugar, la varianza (respuesta Desviación típica) no se ve afectada si al conjunto de valores de la variable se le añade una constante. Si además cada observación es multiplicada por otra constante, en este caso la varianza cambia en relación al cuadrado de la constante (respuesta. La desviación típica cambia en relación al valor absoluto de la constante). Esto queda precisado en la siguiente proposición:

Proposición

Si $Y = ax + b$ entonces $\delta^2 y = a^2 \delta^2 x$

Demostración

Para cada observación x_i de X, $i = 1,...,n$, tenemos una observación de Y que es por definición $Y_i = ax_i + b$. Por la proposición 2.1, se tiene que $\overline{Y} = a\overline{x} + b$. Por tanto, la varianza de Y es

$$\begin{aligned} \mathcal{S}^2{}_Y &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[\underbrace{(a\,x_i + b)}_{y_i} - \underbrace{(a\,\overline{x} + b)}_{\overline{y}}]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} a^2\,(x_i - \overline{x})^2 \\ &= a^2\,\mathcal{S}^2{}_X \end{aligned}$$

Observación

Las consecuencias del anterior resultado eran de esperar: Si los resultados de una medida son trasladados una cantidad b, la dispersión de los mismos no aumenta. Si estos mismos datos se multiplican por una cantidad *a* <1, el resultado tenderá a concentrarse alrededor de su media (menor varianza). Si por el contrario *a*>1 habrá mayor dispersión.

Otra propiedad fundamental de la varianza es la siguiente:

Proposición

Dados *r* grupos, cada uno de ellos formado por n_i observaciones de media $\overline{x_i}$ y de varianza $\delta^2{}_i$. Entonces la varianza, δ^s, del conjunto de todas las $n=n_i+...+n_r$ observaciones vale

$$\mathcal{S}^2 = \frac{1}{n}\sum_{i=1}^{r} n_i\,\mathcal{S}_i^2 + \frac{1}{n}\sum_{i=1}^{r} n_i(\overline{x}_i - \overline{x})^2$$

Demostración

Dicho de otro modo, pretendemos demostrar que la varianza total es igual a la media de las varianzas más la varianza de las medias. Comenzamos denotando

mediante x_{ij} la observación *j*-ésima en el *i*-ésimo grupo, donde *i=1,...,r* y *j= 1,...,*n_1. Entonces

$$
\begin{aligned}
S^2 &= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}[(x_{ij}-\overline{x}_i)+(\overline{x}_i-\overline{x})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 + 2\sum_{i=1}^{r}(\overline{x}_i-\overline{x})\underbrace{\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)}_{=0} + \frac{1}{n}\sum_{i=1}^{r} n_i\,(\overline{x}_i-\overline{x})^2 \\
&\quad \frac{1}{n}\sum_{i=1}^{r} n_i\,S_i^2 + 0 + \frac{1}{n}\sum_{i=1}^{r}(\overline{x}_i-\overline{x})^2
\end{aligned}
$$

Observación

Además de las propiedades que hemos demostrado sobre la varianza (y por tanto sobre la desviación típica), será conveniente tener siempre en mente otras que enunciamos a continuación:

- Ambas son sensibles a la variación de cada una de las puntuaciones, es decir, si una puntuación cambia, cambia con ella la varianza. La razón es que si miramos su definición, la varianza es función de cada una de las puntuaciones.
- Si se calculan a través de los datos agrupados en una tabla, dependen de los intervalos elegidos. Es decir, cometemos cierto error en el cálculo de la varianza cuando los datos han sido resumidos en una tabla estadística mediante intervalos, en lugar de haber sido calculados directamente como datos no agrupados. Este error no será importante si la elección del número de intervalos, amplitud y límites de los mismos ha sido adecuada.
- La desviación típica tiene la propiedad de que en el intervalo

$$(\overline{x}-2S,\overline{x}+2S) \overset{\text{def}}{\sim} \overline{x}\pm 2S$$

se encuentra, al menos, el 75% de las observaciones . Incluso si tenemos muchos datos y estos provienen de una distribución normal (se definirá este concepto más adelante), podremos llegar al 95%.

- No es recomendable el uso de ellas, cuando tampoco lo sea el de la media como medida de tendencia central.

Método abreviado para el cálculo de la varianza

Si una variable *X* toma unos valores para los cuales las operaciones de cálculo de media y varianza son tediosas, podemos realizar los cálculos sobre una variable *Z* definida como

$$Z = \frac{X - x_0}{a}$$

Una vez que han sido calculadas $\overline{X}$ y δ^2_z, obtenemos $\overline{X}$ y δ^2_x teniendo en cuenta que:

$$X = a\,Z + x_0 \Longrightarrow \begin{cases} \overline{x} = a\,\overline{z} + x_0 \\ S_X^2 = a^2 S_Z^2 \end{cases}$$

Grados de libertad

Los *grados de libertad* de un estadístico calculado sobre *n* datos se refieren al número de cantidades independientes que se necesitan en su cálculo, menos el número de restricciones que ligan a las observaciones y el estadístico. Es decir, normalmente *n*-1.

Ilustrémoslo con un ejemplo. Consideramos una serie de valores de una variable,

$$x_i \rightsquigarrow 2, 5, 7, 9, 12$$

que han sido tomados de forma independiente.

Su media es $\overline{X} = 7$ y se ha calculado a partir de las *n*=5 observaciones independientes x_i, que están ligadas a la media por la relación:

$$\overline{x} = \frac{1}{n}\sum x_i$$

Luego el número de grados de libertad de la media es n-1=4.

Si calculamos a continuación la varianza, se han de sumar n cantidades

$$\frac{(x_i - \overline{x})^2}{n}$$

Sin embargo esas cantidades no son totalmente independientes, pues están ligadas por una restricción:

$$\sum_{i=1}^{n}(x_i - \left(\sum_{i=1}^{n} x_i\right)/n) = 0$$

El **número de grados de libertad** del estadístico es el número de observaciones de la variable menos el número de restricciones que verifican, así que en este caso, los grados de libertad de la varianza sobre los *n*=5 datos son también *n*-1 =4.

Un principio general de la teoría matemática nos dice que si pretendemos calcular de modo aproximado la varianza de una población a partir de la varianza de una muestra suya, se tiene que el error cometido es generalmente más pequeño, si en vez de considerar como estimación de la varianza de la población, a la varianza muestral

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

consideramos lo que se denomina **cuasivarianza muestral**, δ^2 que se calcula como la anterior, pero cambiando el denominador por el número de grados de libertad, *n*-1:

$$\widehat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n\,S^2}{n-1}$$

Coeficiente de variación

Hemos visto que las medidas de centralización y dispersión nos dan información sobre una muestra. Nos podemos preguntar si tiene sentido usar estas magnitudes para comparar dos poblaciones. Por ejemplo, si nos piden comparar la dispersión de los pesos de las poblaciones de vacas de dos establos diferentes, δ nos dará información útil.

¿Pero qué ocurre si lo que comparamos es la altura de unas vacas con respecto a su peso? Tanto la media como la desviación típica, $\overline{X}$ y δ , se expresan en las mismas unidades que la variable. Por ejemplo, en la variable altura podemos usar como unidad de longitud el metro y en la variable peso, el kilogramo. Comparar una desviación (con respecto a la media) medida en metros con otra en kilogramos no tiene ningún sentido.

El problema no deriva sólo de que una de las medidas sea de longitud y la otra sea de masa. El mismo problema se plantea si medimos cierta cantidad, por ejemplo la masa, de dos poblaciones, pero con distintas unidades. Este es el caso en que comparamos el peso en kilogramos de una población de 100 vacas con el correspondiente en miligramos de una población de 50 garrapatas.

El problema no se resuelve tomando las mismas escalas para ambas poblaciones. Por ejemplo, se nos puede ocurrir medir a las hormigas con las mismas unidades que las vacas (kilogramos). Si la ingeniería genética no nos sorprende con alguna barbaridad, lo lógico es que la dispersión de la variable peso de las garrapatas sea prácticamente nula (¡Aunque haya algunas que sean 100 veces mayores que otras!)

En los dos primeros casos mencionados anteriormente, el problema viene de la dimensionalidad de las variables, y en el tercero de la diferencia enorme entre las

medias de ambas poblaciones. El coeficiente de variación es lo que nos permite evitar estos problemas, pues elimina la dimensionalidad de las variables y tiene en cuenta la proporción existente entre medias y desviación típica. Se define del siguiente modo:

$$CV = \frac{S_X}{\overline{x}}$$

Basta dar una rápida mirada a la definición del coeficiente de variación, para ver que las siguientes consideraciones deben ser tenidas en cuenta:

- Sólo se debe calcular para variables con todos los valores positivos. Todo índice de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva. De ahí que sólo debemos trabajar con variables positivas, para la que tenemos con seguridad que $\overline{X}\rangle 0$.
- No es invariante ante cambios de origen. Es decir, si a los resultados de una medida le sumamos una cantidad positiva, b>0, para tener $Y=X+b$, entonces, $CV_Y < CV_X$ ya que la desviación típica no es sensible ante cambios de origen, pero si la media. Lo contrario ocurre si restamos (b<0).

$$CV_Y = \frac{S_Y}{\overline{y}} = \frac{S_X}{\overline{x}+b} < \frac{S_X}{\overline{x}} = CV_X$$

- Es invariante a cambios de escala. Si multiplicamos X por una constante a, para obtener $Y = a\,X$, entonces

$$CV_Y = \frac{S_Y}{\overline{y}} = \frac{S_{a\,X}}{a\,\overline{x}} = \frac{\cancel{a}\,S_X}{\cancel{a}\,\overline{x}} = CV_X$$

Observación

Es importante destacar que los coefientes de variación sirven para comparar las variabilidades de dos conjuntos de valores (muestras o poblaciones), mientras que si deseamos comparar a dos individuos de cada uno de esos conjuntos, es necesario usar los valores tipificados.

Ejemplo

Dada las ocasiones que se detecto antibiótico en la leche de un establo de 100 animales, obtener:

1. La variable tipificada *Z*.
2. Valores de la media y varianza de *Z*.
3. Coeficiente de variación de *Z*.

Número de ocasiones que se detecto Antibiótico en la leche de un establo	Num. De vacas
0 -- 4	47
4 -- 10	32
10 -- 20	17
20 -- 40	4
	100

Solución:

Para calcular la variable tipificada

$$Z = \frac{X - \bar{x}}{S_X},$$

partimos de los datos del enunciado. Será necesario calcular en primer lugar la media y desviación típica de la variable original (X= número de ocasiones que se detecta antibiótico en leche).

l_{i-1} -- l_i	x_i	n_i	$x_i\, n_i$	x_i^2 n_i
0 -- 4	2	47	94	188
4 -- 10	7	32	224	1.568
10 -- 20	15	17	255	3.825
20 -- 40	30	4	120	3.600
		n=100	693	9.181

$$\begin{aligned} \overline{x} &= \frac{693}{100} = 6,93 \text{ años} \\ S_X^2 &= \frac{9.181}{100} - 6,93^2 = 43,78 \text{ años al cuadrado} \\ S_X &= \sqrt{43,78} = 6,6 \text{ años} \end{aligned}$$

A partir de estos valores podremos calcular los valores tipificados para las marcas de clase de cada intervalo y construir su distribución de frecuencias:

$$\begin{aligned} z_1 &= \frac{2 - 6,93}{6,6} = -0,745 \\ z_2 &= \frac{7 - 6,93}{6,6} = 0,011 \\ z_3 &= \frac{15 - 6,93}{6,6} = 1,22 \\ z_4 &= \frac{30 - 6,93}{6,6} = 3,486 \end{aligned}$$

z_i	n_i	$z_i\, n_i$	$z_i^2\, n_i$
-0,745	47	-35,015	26,086
0,011	32	0,352	0,004
1,220	17	20,720	25,303
3,486	4	13,944	48,609

	n=100	0,021	100,002

$$\begin{aligned} \bar{z} &= \frac{0,021}{100} \approx 0 \\ S_Z^2 &= \frac{100,02}{100} - 0^2 \approx 1 \\ S_X &= \sqrt{1} = 1 \end{aligned}$$

A pesar de que no se debe calcular el coeficiente de variación sobre variables que presenten valores negativos (y *Z* los presenta), lo calculamos con objeto de ilustrar el porqué:

$$CV = \frac{S_Z}{\bar{z}} = \frac{1}{0} = \infty$$

Es decir, el coeficiente de variación no debe usarse nunca con variables tipificadas.

DISTRIBUCIÓN t DE STUDENT

La mayoría de las veces no se tiene la suerte suficiente como para conocer la varianza de la población de la cual se seleccionan las muestras aleatorias. Para muestras de tamaño n ≥ 30 se proporciona una buena estimación de σ^2 al calcular un valor de S^2.

Si el tamaño muestral es pequeño, los valores de S^2 fluctúan considerablemente de muestra a muestra y la distribución de la variable aleatoria (X - μ) ($S/\sqrt{n}$) se desvía en forma apreciable de una distribución normal estándar. Ahora se está tratando con la distribución de un estadístico que recibe el nombre de T, donde,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

al derivar la distribución muestral de T, se asumirá que la muestra aleatoria se seleccionó de una población normal. Se puede expresar entonces:

$$T = \frac{(\bar{X} - \mu)(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}}$$

donde:

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}}$$

tiene la distribución normal estándar, y

$$V = \frac{(n-1)S}{\sigma^2}$$

Tiene una distribución Ji cuadrada con V = n – 1 grados de libertad. Al muestrear poblaciones normales, puede demostrarse que $\bar{X}$ y S^2 son independientes y en consecuencia lo son Z y V. Ahora ya se está en posibilidad de obtener la distribución T.

La distribución de T se publicó por primera vez en 1908 en un trabajo de W. S. Gosset. En esa época Gosset era empleado de una cervecería irlandesa que desaprobaba la publicación de trabajos de investigación de sus trabajadores. Para evadir esta restricción Gosset publicó su trabajo en secreto con el seudónimo de “Student”, en consecuencia, a la distribución de T usualmente se le llama distribución de t de “Student”, o simplemente distribución de t. Al derivar la

ecuación de esta distribución, Gosset asumió que las muestras se seleccionaban de una población normal. A pesar que esto parecería una suposición restrictiva puede demostrarse que las poblaciones no normales que poseen distribuciones en forma de campana proporcionan valores de T que se aproximan mucho a la distribución de t.

La distribución de T se asemeja a la distribución de Z en que ambas son simétricas alrededor de una media cero. Ambas distribuciones tienen forma de campana, pero la distribución de t varia mas, debido a que el hecho de que los valores de T dependen de los valores de dos cantidades, $\bar{X}$ y S^2, mientras los valores de Z dependen únicamente de los cambios de $\bar{X}$ de muestra a muestra.

La distribución de T difiere de la de Z en que la varianza de la primera depende de la varianza de n y siempre es mayor que 1. Sólo cuando el tamaño muestral es $n \to \infty$ las dos distribuciones son iguales.

La probabilidad de que una muestra aleatoria produzca un valor t = (X - $\mu)/(\sigma/\sqrt{n})$ que caiga entre los dos valores cualesquiera especificados es igual al área bajo la curva de la distribución t entre las dos ordenadas correspondientes a los valores especificados. Seria una tarea tediosa el intentar hacer tablas separadas que dieran las áreas entra cada par de ordenadas posible para todos los valores para n ≤ 30 (Walpole y Myers, 1998)

Distribución t de Student

La distribución t-Student se construye como un cociente entre una normal y la raíz de una X^2 independientes. De modo preciso, llamamos **distribución *t*-Student con *n* grados de libertad**, t_n a la de una v.a. *T*,

$$T = \frac{Z}{\sqrt{\frac{1}{n}\chi^2_n}} \sim t_n$$

donde $Z \rightsquigarrow \mathbf{N}(0,1)$, $\chi_n^2 \rightsquigarrow \chi^2_n$. Este tipo de distribuciones aparece cuando tenemos *n*+1 v.a. independientes

$$X \rightsquigarrow \mathbf{N}(\mu, \sigma^2)$$

$$X_i \rightsquigarrow \mathbf{N}(\mu_i, \sigma_i^2) \qquad i = 1, \ldots, n$$

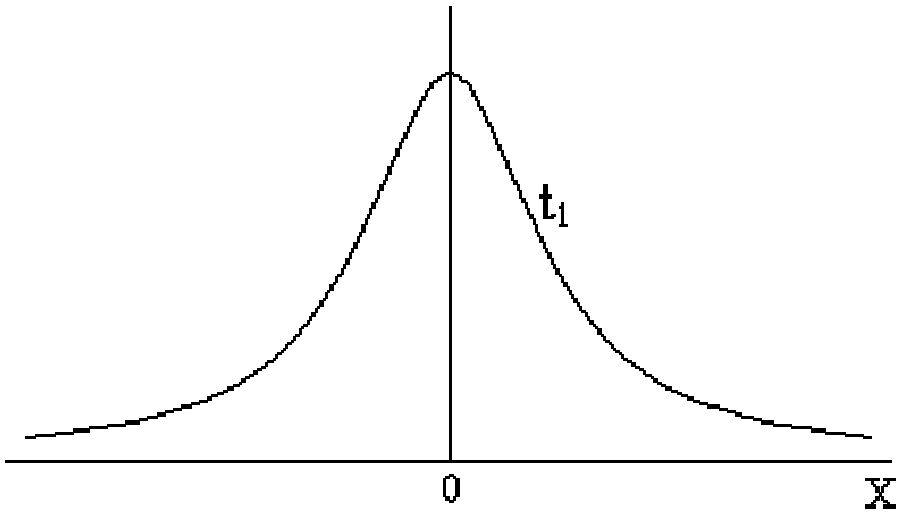
y nos interesa la distribución de

$$T = \frac{\frac{X-\mu}{\sigma}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2}} \rightsquigarrow \mathbf{t}_n$$

La función de densidad de $t_n \rightsquigarrow \mathbf{t}_n$ es

$$f_T(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad \forall t \in \mathbb{R}$$

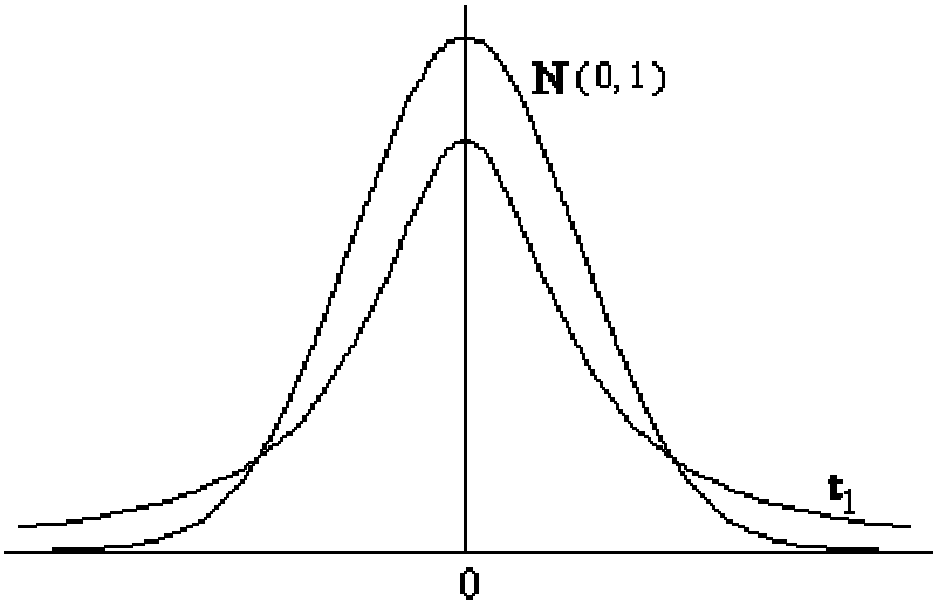
Figura: Función de densidad de una $\mathbf{t}$ de Student



La distribución t de Student tiene propiedades parecidas a N (0,1):

- Es de media cero, y simétrica con respecto a la misma;
- Es algo más dispersa que la normal, pero la varianza decrece hasta 1 cuando el número de grados de libertad aumenta;

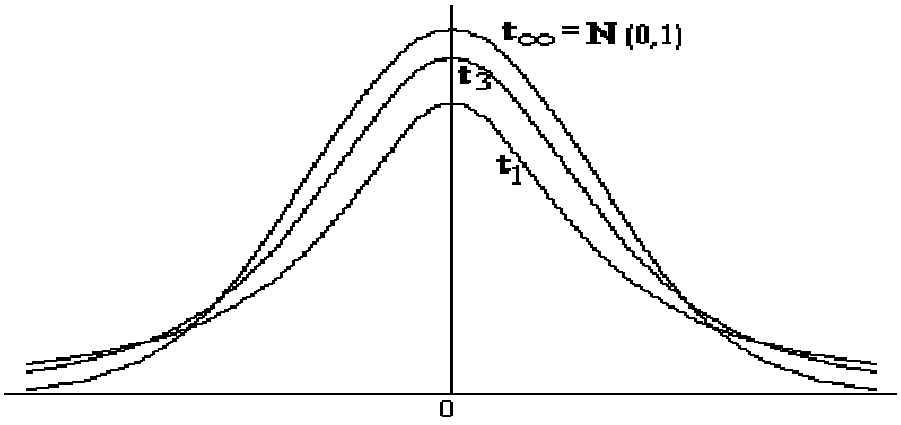
Figura: Comparación entre las funciones de densidad de t_1 y $N(0,1)$.



- Para un número alto de grados de libertad se puede aproximar la distribución de Student por la normal, es decir,

$$t_n \xrightarrow{n\to\infty} N(0,1)$$

Figura: Cuando aumentan los grados de libertad, la distribución de Student se aproxima a la distribución normal tipificada.



- Para calcular

$$\mathcal{P}[T \leq t] = F_T(t) = \int_{-\infty}^{t} f_T(x)\,dx = \int_{-\infty}^{t} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx$$

en lugar de considerar una primitiva de esa función y determinar la integral definida, buscaremos el resultado aproximado en una *tabla de la distribución* t_n.
Intervalo para la media

Como hemos mencionado, los casos anteriores se presentarán poco en la práctica, ya que lo usual es que sobre una población quizás podamos conocer si se distribuye normalmente, pero el valor exacto de los parámetros μ y σ^2 no son conocidos. De ahí nuestro interés en buscar intervalos de confianza para ellos.

El problema que tenemos en este caso es más complicado que el anterior, pues no es tan sencillo eliminar los dos parámetros a la vez. Para ello nos vamos a ayudar de lo siguiente:

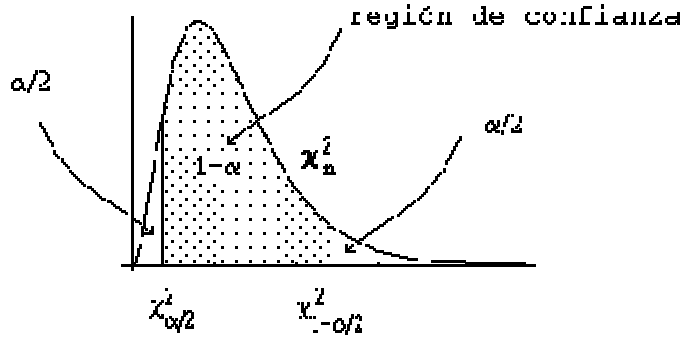
$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathbf{N}(0,1)$$

Por el teorema de Cochran sabemos por otro lado que:

$$\chi^2_{n-1} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

y que además estas dos últimas distribuciones son independientes. A partir de estas relaciones podemos construir una distribución **t** de Student con *n*-1 grados de libertad.

Figura: La distribución t_n es algo diferente a N (0,1) cuando *n* es pequeño, pero conforme éste aumenta, ambas distribuciones se aproximan.



$$T_{n-1} = \frac{Z}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}}$$

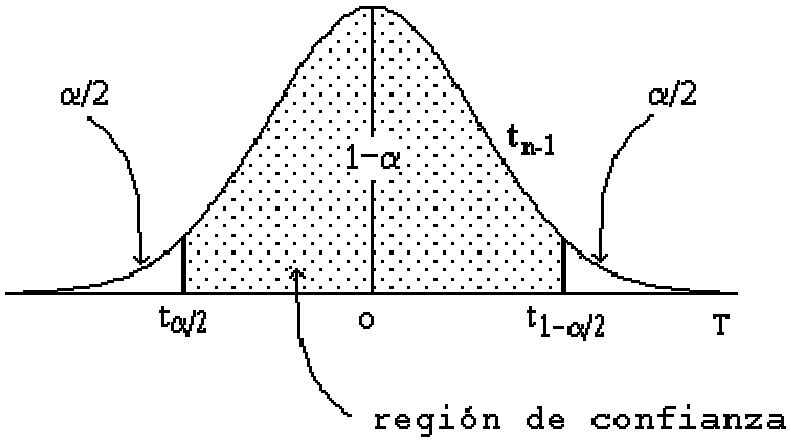
$$= \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}}}$$

Simplificando la expresión anterior tenemos:

$$\boxed{T_{n-1} = \frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}} \rightsquigarrow \mathbf{t}_{n-1}}$$

Dado el nivel de significación 1 - α buscamos en una tabla de $t_n - 1$ el percentil $100 \cdot (1-\alpha/2)$, $t_{n-1,1-\alpha/2}$, el cual deja por encima de si la cantidad $\alpha/2$ de la masa de probabilidad. Por simetría de la distribución de Student se tiene que $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$, luego

Figura: La distribución de Student tiene las mismas propiedades de simetría que la normal tipificada.



$$\begin{cases} \mathcal{P}[T_{n-1} > t_{n-1,1-\alpha/2}] = \frac{\alpha}{2} \\ \mathcal{P}[T_{n-1} < -t_{n-1,1-\alpha/2}] = \frac{\alpha}{2} \end{cases} \iff \mathcal{P}[|T_{n-1}| \leq t_{n-1,1-\alpha/2}] = 1-\alpha$$

El intervalo de confianza se obtiene a partir del siguiente cálculo:

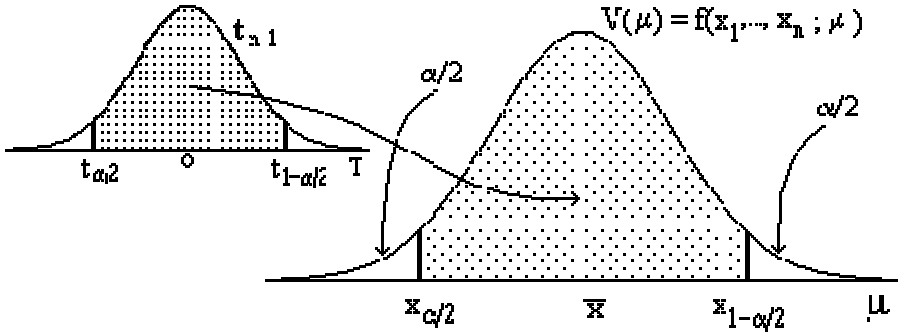
$$|T_{n-1}| \leq t_{n-1,1-\alpha/2} \Rightarrow \frac{|\overline{X}-\mu|}{\hat{S}/\sqrt{n}} \leq t_{n-1,1-\alpha/2}$$

$$\Rightarrow |Xmed-\mu| \leq t_{n-1,1-\alpha/2} \cdot \hat{S}/\sqrt{n}$$

Es decir, el intervalo de confianza al nivel 1 - α para la esperanza de una distribución gaussiana cuando sus parámetros son desconocidos es:

$$\mu = \overline{X} \pm t_{n-1,1-\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Figura: Intervalo de confianza para μ cuando σ^2 es desconocido (caso general).



Al igual que en el caso del cálculo del intervalo de confianza para μ cuando σ^2 es conocido, podemos en el caso σ^2 desconocido, utilizar la función de verosimilitud para representarlo geométricamente. En este caso se usa la notación:

$$x_{\alpha/2} = \overline{x} - t_{n-1,1-\alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}$$

$$x_{1-\alpha/2} = \overline{x} + t_{n-1,1-\alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}$$

Ejemplo

Se requiere encontrar intervalo de confianza a un nivel de significancia de 0.05 para el consumo de alimento promedio en conejos de raza Nueva Zelanda de una granja cunícula. Solo se sabe que el consumo de alimento es una variable aleatoria X de distribución normal, para esto se tiene una muestra de 25 conejos y se obtiene:

$\overline{x}$ = 170 gr

S= 10 gr

Solución:

En primer lugar, en estadística inferencial, los estadísticos para medir la dispersión más convenientes son los insesgados. Por ello vamos a dejar de lado la desviación típica muestral, para utilizar la cuasidesviación típica:

$$S = 10 \Longrightarrow \hat{S} = S\sqrt{\frac{n}{n-1}} = 10\sqrt{\frac{25}{24}} = 10'206$$

Si queremos estimar un intervalo de confianza para μ, es conveniente utilizar el estadístico

$$T = \frac{\overline{x} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \rightsquigarrow \mathbf{t}_{n-1}$$

y tomar como intervalo de confianza aquella región en la que

$$|T| \leq t_{n-1;1-\alpha/2}$$

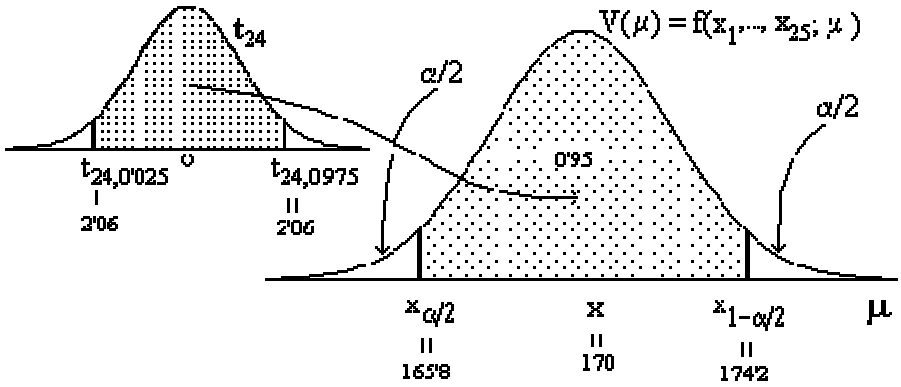
es decir,

$$\left| \frac{170 - \mu}{\frac{10,206}{\sqrt{25}}} \right| \leq t_{24;0,975} = 2,06 \Longrightarrow \mu = 170 \pm 2,06 \cdot \frac{10,206}{5} = 170 \pm 4,204$$

o dicho de forma más precisa: Con un nivel de confianza del 95% podemos decir que la media poblacional está en el intervalo siguiente

$$\mu \in [165,796\,;\,174,204]$$

Figura: Cálculo del intervalo de confianza para la media usando para ello la distribución de Student y la función de verosimilitud asociada, la cual está tiene su máximo en $\bar{x}$, ya que esta estimación puntual de μ es la *máximo verosímil*.



Distribución X^2

En 1900 Karl Pearson propuso el siguiente estadístico de prueba que es una función de los cuadrados de las desviaciones de los números observados con respecto a sus valores esperados, ponderados por el recíproco de sus valores esperados.

$$X^2 = \sum_{i=1}^{k} \frac{[\ddot{}ni - E(ni)]^2}{E(ni)} = \sum_{i=1}^{k} \frac{[ni - npi]^2}{npi}$$

Aunque es una cuestión complicada, se puede demostrar que x^2 tendrá una distribución Ji cuadrada en un muestreo repetitivo, para n grande. Se puede demostrar fácilmente este resultado para el caso de K =2. Si k = 2, entonces $n_2 = n - n_1$ y p1 + p2 = 1. Por lo tanto tiene aproximadamente una distribución normal estándar para n grande. Así para n grande, la X^2 antes indicada, es aproximadamente una variable aleatoria X^2 con un grado de libertad, recordando que el cuadrado de una variable aleatoria normal estándar tiene una distribución X^2.

Existen multitud de situaciones en el ámbito de la salud animal en el que las variables de interés, las cuales no pueden cuantificarse mediante cantidades numéricas, entre las que el investigador esté interesado en determinar posibles relaciones. En este caso tendríamos, a lo sumo, las observaciones agrupadas en forma de frecuencia, dependiendo de las modalidades que presente cada individuo en cada una de las variables, por los que los métodos estudiados en otras ocasiones no serian aplicables.

El objetivo de este tema es el estudio de este tipo de cuestiones en relación con las variables cualitativas (y también v.a. discretas o continuas agrupadas en intervalo). Estos son los contrastes asociados con el estadístico X^2. En general este tipo de tests consisten en tomar una muestra y observar si hay diferencia significativa entre las frecuencias observadas y las especificadas por la ley teórica del modelo que se contrasta, también denominadas ``frecuencias esperadas".

Sin embargo, aunque éste sea el aspecto más conocido, el uso del test X^2 no se limita al estudio de variables cualitativas. Podríamos decir que existen tres

aplicaciones básicas en el uso de este test, y cuyo desarrollo veremos en el transcurso de este capítulo:

Tres son los temas que abordaremos de esta manera:

Test de ajuste de distribuciones:

Es un contraste de significación para saber si los datos de una muestra son conformes a una ley de distribución teórica que sospechamos que es la correcta.

Test de homogeneidad de varias muestras cualitativas:

Sirve para contrastar la igualdad de procedencia de un conjunto de muestras de tipo cualitativo.

Test para tablas de contingencia:

Es un contraste para determinar la dependencia o independencia de caracteres cualitativos.

El estadístico X^2 y su distribución

Sea *X* una v.a. cuyo rango son los valores i = 1, 2, 3.... K , de modo que p_i es la probabilidad de cada valor;

$$X \rightsquigarrow \begin{cases} 1 \rightarrow \mathcal{P}[X=1]=p_1 \\ 2 \rightarrow \mathcal{P}[X=2]=p_2 \\ \dots \\ i \rightarrow \mathcal{P}[X=i]=p_i \\ \dots \\ k \rightarrow \mathcal{P}[X=k]=p_k \end{cases}$$

Este tipo de v.a. puede corresponder a variables ya estudiadas como es el caso de la distribución Binomial

$$X \rightsquigarrow \mathbf{B}(k,p) \Longrightarrow \begin{cases} \text{rango} \equiv 1,2,\ldots,k \\ p_i = \mathcal{P}[X=i] = \binom{k}{i} p^i q^{k-i} \end{cases}$$

pero nosotros vamos a usarla para v.a. más generales. Supongamos que el resultado de un experimento aleatorio es una clase c_1, c_2, ..., c_k(c_i,i = 1...k), que puede representar valores cualitativos, discretos o bien intervalos para variables continuas. Sea p_i la probabilidad de que el resultado del experimento sea la clase c_i. Vamos a considerar contrastes cuyo objetivo es comprobar si ciertos valores p_i^0, propuestos para las cantidades p_i son correctas o no, en función de los resultados experimentales

$$\begin{cases} H_0 : \text{Los } p_i^0 \text{ son correctos} \\ H_1 : \text{Alguno de los } p_i^0 \text{ es falso} \end{cases} \Longleftrightarrow \begin{cases} H_0 : \left| \begin{array}{ll} p_1 = p_1^0 & \text{y} \\ p_2 = p_2^0 & \text{y} \\ \ldots & \\ p_k = p_k^0 & \end{array} \right. \\ H_1 : \left| \begin{array}{ll} p_1 \neq p_1^0 & \text{o bien} \\ p_2 \neq p_2^0 & \text{o bien} \\ \ldots & \\ p_k \neq p_k^0. & \end{array} \right. \end{cases}$$

Mediante muestreo aleatorio simple, se toma una muestra de tamaño *n* y se obtienen a partir de ella unas *frecuencias observadas* de cada clase que representamos mediante O_1, O_1, ..., O_k

Clase	Frec. Abs.
c_i	O_i
c_1	O_1
c_2	O_2
...	...
c_k	O_k
	$\sum_{i=1}^{k} O_i = n$

Supongamos que la hipótesis nula es cierta. Al ser $p_i=p_i^0$ la proporción de elementos de la clase c_i en la población, el número de individuos de que presentan esta modalidad al tomar una muestra de tamaño n, es una v.a. de distribución binomial, B (n, Pi^o). Por tanto la *frecuencia esperada* de individuos de esa clase es

$$\mathcal{E}_i = n \cdot p_i^0 \qquad \forall i = 1, 2, \ldots, k$$

$$\sum_{i=1}^{k} \mathcal{E}_i = n \cdot \sum_{i=1}^{k} p_i^0 = n$$

Obsérvese que a diferencia de las cantidades O_i, que son las frecuencias que realmente se obtienen en una muestra, las frecuencias esperadas no tienen por que ser números enteros. De cualquier modo, bajo la suposición de que H_0 es cierta cabe esperar que las diferencias entre las cantidades $\in_i$ y O_i sea pequeña.

Pearson propuso el estadístico

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - \mathcal{E}_i)^2}{\mathcal{E}_i}$$

El cual, siguiendo la línea de razonamiento anterior debe tomar valores pequeños si H_0 es cierta. Si al tomar una muestra, su valor es grande eso pone en evidencia que la hipótesis inicial es *probablemente* falsa. Para decidir cuando los valores de X^2 son grandes es necesario conocer su ley de probabilidad. Se tiene entonces el siguiente resultado

Teorema

Ley asintótica para X^2 , so al hipótesis es cierta, entonces X^2 se distribuye aproximadamente como

$$X^2 = \sum_{i=1}^{k} \frac{(Oi - \in i)^2}{\in i} \xrightarrow{=} X^2_{k-p-h}$$

Donde el numero de grados de libertad depende de:

El numero de *K,* de clases usadas;

El numero de parámetros estimados a partir de la muestra para calcular los $\in i$. Por ejemplo si todas las cantidades p_i^0 son especificadas desde p=o.

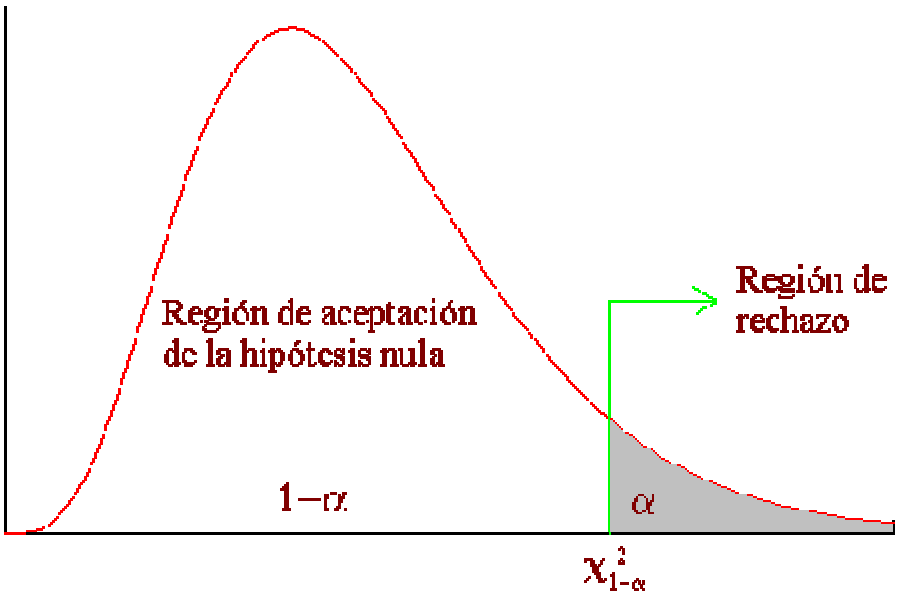
El numero de relaciones o condiciones impuestas a los $\in i.$ Por ejemplo si la única condición sobre los $\in i.$ es: $\sum_{i=1}^{k} \in i = 1$ entonces *h=1.*

La aproximación mejora cuando n es grande y los *pi* son cercanos a $\frac{1}{2}$.

Como sólo son los valores grandes de X^2 los que nos llevan a rechazar H_0, la región crítica es:

$$C = (\chi^2_{k-p-h,1-\alpha}, \infty)$$

Figura: Región crítica (sombreada) para un contraste con el estadístico X^2.



es decir,

$$\text{sean} \begin{cases} \chi^2_{exp} = \sum_{i=1}^{k} \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i} \\ \chi^2_{teo} = \chi^2_{k-p-h,1-\alpha} \end{cases} \longrightarrow \begin{cases} \text{Si } \chi^2_{exp} \leq \chi^2_{teo} \text{ no rechazamos } H_0 \text{ ;} \\ \text{Si } \chi^2_{exp} > \chi^2_{teo} \text{ se rechaza } H_0 \text{ y se acepta } H_1 \text{ .} \end{cases}$$

Observación

A pesar de que el contraste parece ser bilateral al ver la expresión de la relación, la forma de $\mathcal{C}$, nos indica que el contraste es unilateral: Sólo podemos saber si existe desajuste entre los esperado y lo observado, pero no podemos contrastar hipótesis alternativas del tipo ``p_i mayor que cierto valor''.

Observación

Obsérvese que en realidad X^2 no es una variable aleatoria continua: Los posibles resultados de la muestra se resumen en las cantidades O_1, O_2, ..., O_k, que *únicamente* toman valores discretos. Luego las cantidades

$$\chi^2_{exp}(\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_k)$$

sólo puede tomar un número finito de valores distintos (aunque sean cantidades con decimales). Por tanto su distribución *no es continua*. Luego al realizar la aproximación mencionada hay que precisar en qué condiciones el error cometido es pequeño. De modo aproximado podemos enunciar el siguiente criterio que recuerda al de la aproximación binomial por la distribución normal:

1.

n>30;

2.

$\mathcal{E}_i = n \cdot p_i > 5$ para todo $i = 1, \ldots, k$.

Sin embargo esta regla resulta demasiado estricta a la hora de aplicarla en la práctica. Se utiliza entonces una regla más flexible y que no sacrifica demasiada precisión con respecto a la anterior:

1. Para ninguna clase ocurre que $\mathcal{E}_i = n \cdot p_i < 1$
2. $\mathcal{E}_i = n \cdot p_i > 5$ para casi todos los i= 1.... K., salvo a lo sumo un 20% de ellos.

Si a pesar de todo, estas condiciones no son verificadas, es necesario agrupar las clases que tengan menos elementos con sus adyacentes.

Observación

El lector puede considerar los contrastes con el estadístico X^2 como una generalización del contraste de proporciones. Para ello le invitamos a estudiar el siguiente ejemplo.

Ejemplo

Se desea saber si cierta enfermedad afecta del mismo modo a los cabritos machos que a las hembras. Para ello se considera una muestra de *n* = 618 individuos que padecen la enfermedad, y se observa que 341 son machos y el resto son hembras. ¿Qué conclusiones se obtiene de ello?

Solución:

El contraste a realizar se puede plantear de dos formas que después veremos que son equivalentes:

Contraste de una proporción:

Si *p* es el porcentaje de machos en la población de enfermos, podemos considerar el contraste:

$$\begin{cases} H_0 : p = 1/2 \\ H_1 : p \neq 1/2 \end{cases}$$

De la muestra obtenemos la siguiente estimación puntual del porcentaje de enfermos de machos:

$$\hat{p} = 341/618 = 0,55178$$

Para ver si esto es un valor ``coherente'' con la hipótesis nula, calculemos la significatividad del contraste:

$$Z_{exp} = \frac{\hat{p} - p}{\sqrt{p * q/n}} \rightsquigarrow \mathbf{N}(0,1).$$

Por otro lado,

$$Z_{exp} = \frac{0,55178 - 0,5}{\sqrt{0,5 \times 0,5/60}} = 2,574$$

Como el contraste es de tipo bilateral, la significatividad del contraste es (buscando en la tabla de la distribución normal):

$$\mathcal{P}[|Z| > 2,574] = 2 \cdot \mathcal{P}[Z > 2,574] = 2 * 0,005 = 1\% < 5\%$$

Lo que nos indica que se ha de rechazar la hipótesis nula y aceptar la hipótesis alternativa, es decir, afirmamos que existe una evidencia significativa a favor de la hipótesis de que la enfermedad no afecta por igual a machos y hembras.

Contraste con el estadístico X^2:

En este caso planteamos el contraste:

Ho :	p machos = ½ y p hembras = ½
H1	p hembras ≠ ½ o bien P hembras ≠ ½

Para resolverlo escribimos en una tabla los frecuencias muestrales observadas de machos y hembras, junto a los valores esperados en el caso de que la hipótesis nula fuese cierta:

	frecuencias	frecuencias		
	observadas	esperadas	diferencia	
	O_i	$\mathcal{E}_i$	$O_i - \mathcal{E}_i$	$(O_i - \mathcal{E}_i)^2/\mathcal{E}_i$
Machos	341	$618 \times 1/2 = 309$	9	$32^2/309$
Hembras	277	$618 \times 1/2 = 309$	-9	$(-32)^2/309$
	618	618	0	6,63

Consideremos entonces el estadístico

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - \mathcal{E}_i)^2}{\mathcal{E}_i} \overset{\approx}{\rightsquigarrow} \chi^2_{k-p-h} = \chi^2_{2-0-1} = \chi^2_1$$

donde:

- *k*=2es el numero de modalidades posibles que toma la variable sexo: *machos* y *hembras*;

- *p*=0 es el número de parámetros estimados;
- *h*=1 es el números de restricciones impuestas a los valores esperados. Sólo hay una (que es habitual), que consiste en que el número esperado de enfermos entre machos y hembras es 60.

El estadístico calculado sobre la muestra ofrece el valor experimental:

$$\chi^2_{exp} = 6,63$$

que es el percentil 99 de la distribución X_1^2 . De nuevo se obtiene que la significatividad del contraste es del 1%<5%.

En conclusión, con los dos métodos llegamos a que hay una fuerte evidencia en contra de que hay el mismo porcentaje de machos y hembras que padecen la enfermedad. La ventaja de la última forma de plantear el contraste (diferencia entre frecuencias observadas y esperadas) es que la técnica se puede aplicar a casos más generales que variables dicotómicas.

Observación

Hay una fórmula alternativa para el cálculo de X^2 cuya expresión es más fácil de utilizar cuando realizamos cálculos:

Proposición

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n$$

Demostración

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{k} \frac{(O_i - \mathcal{E}_i)^2}{\mathcal{E}_i} \\
&= \sum_{i=1}^{k} \frac{O_i^2 - 2\,O_i\mathcal{E}_i + \mathcal{E}_i^2}{\mathcal{E}_i} \\
&= \sum_{i=1}^{k} \frac{O_i^2}{\mathcal{E}_i} - 2\sum_{i=1}^{k} O_i + \sum_{i=1}^{k} \mathcal{E}_i \\
&= \sum_{i=1}^{k} \frac{O_i^2}{\mathcal{E}_i} - 2n + n \\
&= \sum_{i=1}^{k} \frac{O_i^2}{\mathcal{E}_i} - n
\end{aligned}
$$

Distribución X^2 con un grado de libertad

Si consideramos una v.a. $Z \rightsquigarrow \mathbf{N}(0,1)$, la v.a. $X=Z^2$ se distribuye según una ley de probabilidad **distribución X^2 con un grado de libertad**, lo que se representa como

$$X \rightsquigarrow \chi_1^2$$

Si tenemos *n* v.a. independientes, la suma de sus cuadrados respectivos es una distribución que denominaremos **ley de distribución X^2 con *n* grados de libertad**, X_n^2.

$$\{Z_i\}_{i=1}^{n} \rightsquigarrow \mathbf{N}(0,1) \implies \sum_{i=1}^{n} Z_i^2 \rightsquigarrow \chi_n^2$$

La media y varianza de esta variable son respectivamente:

$$\begin{aligned} \mathbf{E}[X] &= n \\ \mathbf{Var}[X] &= 2n \end{aligned}$$

y su función de densidad es:

$$f_{\chi^2_n}(x) = \begin{cases} 0 & \text{si } x \in (-\infty, 0] \\ \dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{si } x \in (0, \infty) \end{cases}$$

Los percentiles de esta distribución que aparecen con más frecuencia en la práctica los podemos encontrar en la tabla 5.

Figura: Función de densidad de X_n^2 para valores pequeños de *n*.

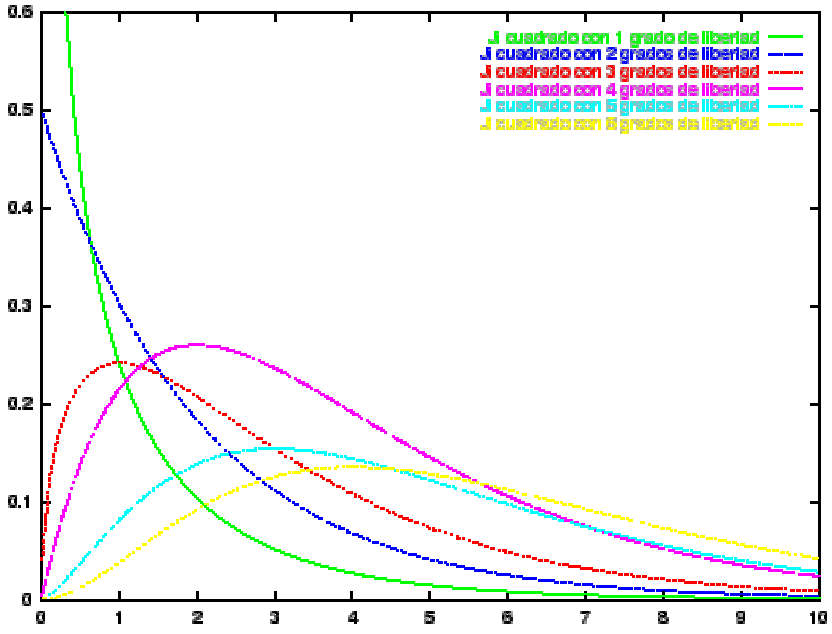
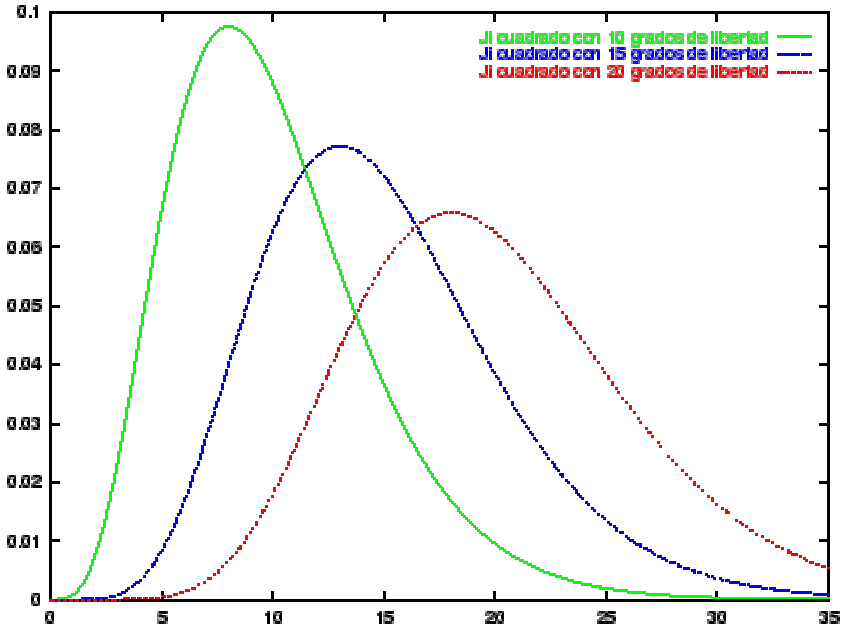


Figura: Función de densidad de X_n^2 para valores grandes de *n*.



En consecuencia, si tenemos X1..... Xn, v.a. independientes, donde cada $X_i \sim \mathbf{N}(\mu_i, \sigma_i^2)$, se tiene

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \sim \chi_n^2$$

Observación

La ley de distribución X^2 muestra su importancia cuando queremos determinar la variabilidad (sin signo) de cantidades que se distribuyen en torno a un valor central siguiendo un mecanismo normal. Como ilustración tenemos el siguiente ejemplo:

Ejemplo

Un instrumento para medir el nivel de inmunoglobulinas en calostro, ofrece resultados bastantes aproximados con la realidad, aunque existe cierta cantidad de error ϵque se distribuye de modo normal con media 0 y desviación típica $\sigma = 2$.

$$X_{\text{real}} = X_{\text{exp}} + \epsilon, \qquad \epsilon \rightsquigarrow \mathbf{N}\left(\mu = 0, \sigma^2 = 2^2\right)$$

Se realizan mediciones de los niveles de inmunoglobulinas dados por el instrumento en un grupo de *n*=100 calostro extraído de 100 vacas. Nos interesa medir la cantidad de error que se acumula en las mediciones de todas las vacas. Podemos plantear varias estrategias para medir los errores acumulados. Entre ellas destacamos las siguientes:

1. Definimos el error acumulado en las mediciones de todas las vacas como

$$E_1 = \sum_{i=1}^{n} \epsilon_i$$

¿Cuál es el valor esperado para E_1?

2. Definimos el error acumulado como la suma de los cuadrados de todos los errores (cantidades positivas):

$$E_2 = \sum_{i=1}^{n} \epsilon_i^2$$

¿Cuál es el valor esperado para E_2?

A la vista de los resultados, cuál de las dos cantidades, E_1 y E_2, le parece más conveniente utilizar en una estimación del error cometido por un instrumento.

Solución:

Suponiendo que todas las mediciones son independientes, se tiene que

$$E_1 = \sum_{i=1}^{n} \epsilon_i = \underbrace{\underbrace{\epsilon_1}_{N(\mu,\sigma^2)} + \underbrace{\epsilon_2}_{N(\mu,\sigma^2)} + \cdots + \underbrace{\epsilon_n}_{N(\mu,\sigma^2)}}_{N(\mu, n\cdot\sigma^2)} \Longrightarrow \mathbf{E}[E_1] = \mu = 0$$

De este modo, el valor esperado para E_1 es 0, es decir, que los errores e_i van a tender a compensarse entre unas vacas y otras. Obsérvese que si μ no fuese conocido a priori, podríamos utilizar E_1, para obtener una aproximación de μ

$$\mu \approx \frac{E_1}{n}$$

Sin embargo, el resultado E_1 no nos indica en qué medida hay mayor o menor dispersión en los errores con respecto al 0. En cuanto a E_2 podemos afirmar lo siguiente:

$$E_2 = \sum_{i=1}^{n} \epsilon_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{\epsilon_i}{\sigma}\right)^2 = \sigma^2 \underbrace{[\underbrace{\left(\frac{\epsilon_1}{\sigma}\right)^2}_{\chi^2_1} + \cdots + \underbrace{\left(\frac{\epsilon_n}{\sigma}\right)^2}_{\chi^2_1}]}_{\chi^2_n} \Longrightarrow \mathbf{E}[E_2] = n\cdot\sigma^2 = 400$$

En este caso los errores no se compensan entre sí, y si σ^2 no fuese conocido, podría ser ``estimado'' de modo aproximado mediante

$$\sigma^2 \approx \frac{E_2}{n}$$

Sin embargo, no obtenemos ninguna información con respecto a μ.

En conclusión, E_1 podría ser utilizado para calcular de modo aproximado μ, y E_2 para calcular de modo aproximado σ^2. Las dos cantidades tienen interés, y ninguna lo tiene más que la otra, pues ambas formas de medir el error nos aportan información.

El siguiente resultado será de importancia más adelante. Nos afirma que la media de distribuciones normales independientes es normal pero *con menor varianza* y relaciona los grados de libertad de una v.a. con distribución X^2, con los de un estadístico como la varianza.

LITERATURA REVISADA

Bioestadística: Métodos y Aplicaciones. Universidad de Málaga. **http://www.bioestadistica.uma.es/libro/**

Castillo P.J, J.G.Arias. 1998. Estadística inferencial básica. Grupo editorial Ibero América. México.

Cochran. G. William. 1980. Diseños experimentales. Editorial Trillas. México

Infante G. S. 1997. Métodos estadísticos. Editorial Trillas. México.

Kreyszig Erwin. 1979. Introducción a la estadística matemática. Editorial Limusa. México

Montgomery D. C.1991.Diseño y análisis de experimentos. Grupo editorial Iberoamérica. México.

Ostle B; Estadística aplicada Ed. Limusa 1983 México.

Mendenhall W. Estadística Matemática con aplicaciones Ed. Iberoamaricana 1986. México

Walpole E.R; Myers H.R. Probabilidad y Estadística 1998. Ed. Limusa, México

LITERATURA REVISADA

Bioestadística: Métodos y Aplicaciones. Universidad de Málaga. http://www.bioestadistica.uma.es/libro/

Castillo P.J, J.G.Arias. 1998. Estadística inferencial básica. Grupo editorial Ibero América. México.

Cochran. G. William. 1980. Diseños experimentales. Editorial Trillas. México

Infante G. S. 1997. Métodos estadísticos. Editorial Trillas. México.

Kreyszig Erwin. 1979. Introducción a la estadística matemática. Editorial Limusa. México

Montgomery D. C.1991.Diseño y análisis de experimentos. Grupo editorial Iberoamérica. México.

Ostle, B. 1965. Estadística aplicada. Primera edición. Editorial Limusa. México.

Rodríguez del A. J. 1991. Métodos de investigación pecuaria. Editorial Trillas. México.

Snedecor W. George, W. G. Cochran. 1979 métodos estadísticos. Editorial Continental. México.

Steel G.D Robert, J. H. Torrie.1981. 2ª. Principles and procedures of statistics a biometrical approach. 2ª. Ed. Editorial Mc Graw-Hill. USA.

Walpole, R. E.1992. Probabilidad y estadística. Cuarta Edición. Editorial Mc GrawHill. México.